

Sommersemester 2024

PubMedMA: Erstellung einer Datenbank für Meta-Analysen mithilfe eines Large Language Models und PubMed

Aschlusspaper

zur Abgabe im Modul

Forschungsprojekt B

im Studiengang Business Application Architecture
an der Fakultät Wirtschaftsinformatik
der Hochschule Furtwangen University

vorgelegt von

Niko Kauz
nikokauz@icloud.com
Salzstrasse 2
78073 Bad Dürrheim

Betreuer:

Prof. Dr. Peter Schanbacher

Eingereicht am

Datum: 26.07.2024

Eidesstattliche Erklärung

Ich versichere, dass ich die vorstehende Arbeit selbstständig und ohne unzulässige fremde Hilfe verfasst und hierzu keine anderen als die angegebenen Hilfsmittel verwendet habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus fremden Quellen entnommen wurden, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt oder an anderer Stelle veröffentlicht.

Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben kann.

Bad Dürrheim, den 26. Juli 2024

A handwritten signature in black ink, appearing to read "N. Kauz".

Niko Kauz

Zusammenfassung

In der vorliegenden Arbeit wurde eine Datenbank, PubMedMA (PubMed Meta-Analyses), entwickelt, die auf der Analyse biomedizinischer Forschungsarbeiten und Zusammenfassungen durch ein Large Language Model (LLaMa3) basiert. Ziel ist es, medizinische Daten in einer strukturierten Datenbank als Medical Records zu sammeln und für Meta-Analysen nutzbar zu machen. Die Implementierung nutzt Technologien wie Python, Jupyter-Notebook, PostgreSQL und Docker. Durch den Einsatz von Large Language Models wird die Analyse und Extraktion von Daten aus der PubMed-Datenbank für medizinische Paper und Zusammenfassungen als Pipeline implementiert und automatisiert. Die generierten Medical Records umfassen Informationen zu Krankheiten, Medikamenten, medizinischen Gruppen und Ergebnissen sowie Resultaten und werden in der PubMedMA-Datenbank gespeichert. Diese ermöglicht umfassende medizinische Meta-Analysen und fördert die Forschung durch einfache Zugänglichkeit sowie strukturierte Abfragemöglichkeiten, die einen übersichtlichen Zugriff auf die Daten gewährleisten. Durch die Evaluation der Ergebnisse konnte eine meist gute Genauigkeit der Datenextraktion und -verarbeitung bestätigt werden. Die Datenbank erweist sich als potenziell wertvolles Tool für medizinische Forscher und die Pharmaindustrie zur Beschleunigung von Forschungs- und Entwicklungsprozessen und zur Verbesserung von Behandlungsmethoden. Weiter bildet es eine solide Grundlage für zukünftige Entwicklungen, indem mehr Daten, Feedbackmechanismen und externe Datenbanken eingebaut werden.

Schlüsselwörter: Meta-Analysen, Large Language Models (LLM), Medizinische Paper, PubMed, Natural Language Processing (NLP), Biomedizinische Forschung, Automatisierte Datenextraktion

Abstract

In the present work, a database, PubMedMA (PubMed Meta-Analyses), was developed based on the analysis of biomedical research papers and summaries using a Large Language Model (LLaMa3). The aim is to collect medical data in a structured database as medical records and make them usable for meta-analyses. The implementation uses technologies such as Python, Jupyter-Notebook, Post-greSQL and Docker. By using Large Language Models, the analysis and extraction of data from the PubMed database for medical papers and abstracts is implemented and automated as a pipeline. The generated medical records include information on diseases, drugs, medical groups and outcomes as well as results and are stored in the PubMedMA database. This enables comprehensive medical meta-analyses and promotes research through simple accessibility and structured query options that ensure clear access to the data. Evaluation of the results has confirmed that the accuracy of data extraction and processing is generally good. The database proves to be a potentially valuable tool for medical researchers and the pharmaceutical industry to accelerate research and development processes and improve treatment methods. It also provides a solid foundation for future developments by incorporating more data, feedback mechanisms and external databases.

Keywords: Meta-analyses, Large Language Models (LLM), medical papers, PubMed, Natural Language Processing (NLP), biomedical research, automated data extraction

Inhaltsverzeichnis

1	<i>Einleitung</i>	10
2	<i>Methodik</i>	11
2.1	Implementierungs-Prozess	11
2.2	Umgebung, Technologien und Werkzeuge	12
2.3	Data Collection	13
2.4	Die Datenquelle	13
2.5	API zu PubMed und NCBI	13
2.6	Vorstellung der Daten	15
2.7	Die Sammlung der Daten	16
2.8	Data Preparation	18
2.9	Persistierung der Daten	18
2.10	Anmerkungen	18
3	<i>Prompt Engineering</i>	19
3.1	Prompt zum Filtern von Paper	19
3.2	Prompt zur Bestimmung der Anzahl an Gruppen	20
3.3	Prompt zum Generieren von Medical Records	20
3.4	Prompt zur Evaluation von Medical Records	21
4	<i>Analyse-Prozess</i>	22
4.1	Automatisierung durch eine Pipeline	24
4.2	Auswirkungen von Prompt Änderungen	25
4.3	Weitere Anmerkungen	25
5	<i>Post Processing</i>	26
5.1	Interaktionen mit der Datenbank	26
5.2	Erweiterte Pipeline: Vom LLM zur Datenbank	27
5.3	Generierte Records und ersten Abfragen	27
6	<i>Evaluation</i>	28
6.1	Evaluationskonzept	28
6.2	Umsetzung der Evaluation	29
6.3	Evaluations-Algorithmen	29
6.4	Bewertung der Evaluation	32
6.4.1	Accuracy der korrekt klassifizierten Paper	32
6.4.2	Accuracy der Medical Records	32
6.4.3	Genauigkeit der Records	33
6.5	Abschließende Bewertung	37

7 Ergebnisse und Diskussion	38
7.1 Einhaltung der Enumeration.....	38
7.1.1 Outcome Category	38
7.1.2 Enumeration – Outcome Evaluation	38
7.2 Die Records	40
7.2.1 Doppelte Records	40
7.2.2 Anzahl der Records pro Gruppe.....	41
7.2.3 Records für ein Medikament	42
7.2.4 Records zum Vergleich von zwei verschiedenen Medikamenten	44
7.2.5 Records nach Jahr und Medikament.....	45
7.2.6 Schwankende und schwindende Records	45
7.2.7 Anmerkungen zu den Records.....	46
7.3 Bewertung der Datenbank	47
8 Fazit und Ausblick.....	48
A. Data Collection	50
B. Prompt Engineering	54
B.1 Filter Prompt	54
B.2 Group Prompt.....	55
B.3 Records Prompt	59
B.4 Evaluation Prompt.....	69
C. Analyse.....	71
D. Ergebnisse und Diskussion	73
D.1 Alte Daten.....	73
D.2 Weitere Tabellen und Daten	74
D.3 Vergleich LLaMa3.1 und LLaMa3	76

Abbildungsverzeichnis

Abbildung 2-1: Visualisierter High-Level Prozess zur Umsetzung des Projektes ..	11
Abbildung 2-2: Prozess der Datensammlung	17
Abbildung 4-1: Analysierte Paper, Dauer und fehlgeschlagene Läufe bei einem Komplett durchlauf des Algorithmus auf den gesamten Datensatz.....	24
Abbildung 5-1: Teilausschnitt der 3.659 Medical Records in der DB-Tabelle	27
Abbildung 6-1: Prozess zur Evaluation der LLaMa3-Analyse und dessen Antworten	29
Abbildung 6-2: Beispiel Dictionary mit Hilfsvariablen für die Evaluation	30
Abbildung 6-3: Ergebnis der Accuracy für die Anzahl der korrekt geflaggten Paper	32
Abbildung 6-4: Ergebnisse der Accuracy für die Evaluation der Anzahl an Records	33
Abbildung 6-5: Ähnlichkeitswerte der Attribute zwischen einem realen und vorhergesagten Record.....	34
Abbildung 6-6: Balkendiagramm zur Visualisierung der Accuracy bei den passenden Thresholds	36
Abbildung 7-1: Anzahl der Records nach Outcome Category für eine Teilmenge	39
Abbildung 7-2: Anzahl der Records nach Outcome Evaluation.....	39
Abbildung 7-3: Schnelle Übersicht der strukturierten Medical Records Daten in der PubMedMA-DB	40
Abbildung 7-4: Duplikat eines Medical Records mit unterschiedlich repräsentierten Resultaten	41
Abbildung 7-5: Anzahl der Medical Records nach Medikament.....	41
Abbildung 7-6: Anzahl der Medical Records nach Medikament.....	41
Abbildung 7-7: Anzahl der Paper nach Krankheit.....	42
Abbildung 7-8: Die Medical Records zu Migraine und Erenumab.....	42
Abbildung 7-9: Medical Records zu Erenumab.....	43
Abbildung 7-10: Betrachtung des HIT-6 Score und des Outcomes für verschiedene Medikamente.....	43
Abbildung 7-11: Betrachtung des MIDAS Score und des Outcomes für verschiedene Medikamente.....	43
Abbildung 7-12: Vergleich zwischen zwei Medikamenten für eine Krankheit und ein Medical Outcome.....	44
Abbildung 7-13: Vergleich von zwei Medikamenten in den Outcome Kategorien und Evaluationen.....	44
Abbildung 7-14: Anzahl der veröffentlichten Paper nach Jahr für Erenumab	45

Abbildung 7-15: Anzahl der veröffentlichten Paper nach Jahr für Lasmiditan	45
Abbildung 7-16: Erkennung von komplexen Records, die nicht bei jedem Lauf erkannt werden	46
Abbildung A-1: Aufbau eines Papers bzw. Abstracts in PubMed.....	50
Abbildung A-2: Echte Daten des Papers mit der PubMed ID 34743579	51
Abbildung A-3: Beschreibung des master-DF.....	52
Abbildung A-4: Aufbau des master-DF mit Daten von einigen Paper.....	53
Abbildung C-1: Begründungen von LLaMa3 für das Nichtfinden von Records in einem synthetischen Paper	71
Abbildung C-2: Kurzer Durchlauf mit einem kombinierten Prompt. Dieser dauerte nur knapp 4 Stunden 30 Minuten statt 12 Stunden	71
Abbildung C-3: Stabilität der Pipeline mit wenigen Fehlern.....	72
Abbildung D-1: Übersicht der Daten in einer alten Version der DB	73
Abbildung D-2: Records ohne richtiges Resultat	74
Abbildung D-3: Paper ohne numerische Resultate	74
Abbildung D-4: Anzahl der veröffentlichten Paper pro Jahr	75
Abbildung D-5: Anzahl der Records nach Outcome Category und Evaluation	75

Tabellenverzeichnis

Tabelle 2-1: Die Daten und Attribute eines Papers in PubMed.....	15
Tabelle 6-1: Accuracy für die einzelnen Attribute bei verschiedenen Thresholds (Alle Angaben in Prozent)	35
Tabelle 7-1: Alte Records für HIT-6 Scores für das Medikament Erenumab.....	46
Tabelle 7-2: Alte Records für MIDAS Scores für das Medikament Erenumab.....	46
Tabelle D-1: Zeilen mit schwer verstehbaren Daten	73
Tabelle D-2: Zeile mit leicht verstehbaren Daten.....	73

Abkürzungsverzeichnis

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
DB	Datenbank
DF	DataFrame
JSON	JavaScript Object Notation
KI	Künstliche Intelligenz
LLM	Large Language Model
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing

1 Einleitung

Die fortschreitende Digitalisierung und der Einsatz von Künstlicher Intelligenz (KI) sowie Natural Language Processing (NLP) haben das Potenzial die biomedizinische Forschung und das Gesundheitswesen zu revolutionieren. Das National Center for Biotechnology Information (NCBI) bietet durch PubMed eine umfassende Datenbank (DB) von Peer-reviewed und veröffentlichten Paper, die eine unschätzbare Ressource für Wissenschaftler und Ärzte darstellen. In diesem Paper wird die Nutzung eines Large Language Models (LLM) zur Analyse von PubMed-Papers beschrieben. Ziel ist es, medizinische Datensätze in Form von Medical Records über mehrere Prompts automatisiert in einer Pipeline zu generieren und diese in einer strukturierten DB zu speichern, um Meta-Analysen zu ermöglichen.

Die Erstellung einer solchen DB, hier PubMedMA-DB genannt, bietet zahlreiche Vorteile für die medizinische Forschung und Praxis. Eine zentralisierte und gut strukturierte Sammlung von Studien zu Krankheiten, Medikamenten und deren Outcomes ermöglicht eine effiziente und gezielte Recherche. Forscher können durch Meta-Analysen umfassendere und robustere Erkenntnisse gewinnen, was letztlich zu besseren Behandlungsmethoden und therapeutischen Strategien führt. Zudem können klinische Entscheidungen durch den Zugang zu aggregierten Daten und Vergleichsstudien verbessert werden, was die Patientensicherheit und -versorgung erhöht.

Ein besonderer Vorteil ergibt sich für Unternehmen und die Pharmaindustrie. Diese können die DB nutzen, um schneller und gezielter auf relevante Studien und Ergebnisse zuzugreifen, was die Forschung und Entwicklung neuer Medikamente beschleunigen kann. Durch die systematische Übersicht und Analyse der vorhandenen Daten können Trends und Lücken in der aktuellen Forschung identifiziert werden, wodurch Innovationspotenziale erkannt und genutzt werden können. Derzeit fehlt in der medizinischen Forschung eine solche umfassende und zugängliche Übersicht, was die Effizienz und Effektivität von Forschungsprozessen hemmt. Die Implementierung dieser DB könnte somit einen entscheidenden Beitrag zur Optimierung und Beschleunigung biomedizinischer Forschung leisten.

Dieses Projekt zielt darauf ab, die umfangreichen Informationen aus PubMed-Papers systematisch als Medical Records zu extrahieren und in einer DB zu speichern, die spezifisch für Meta-Analysen entwickelt wurde. Dabei handelt es sich bei den medizinischen Records um Studien zu Krankheiten, Medikamenten, Therapien, medizinischen Gruppen und deren Outcomes sowie Resultate. Durch den Einsatz eines LLM wird die Extraktion und Strukturierung der Daten automatisiert, was eine schnelle und genaue Analyse großer Datenmengen ermöglicht. In diesem Projekt wird diese Aufgabe als komplexes NLP Problem betrachtet und mit entsprechenden Werkzeugen und Modellen behandelt.

2 Methodik

In diesem Kapitel wird der Implementierungsprozess des Projekts detailliert beschrieben. Es werden die einzelnen Schritte erläutert, die zur Realisierung des Projekts notwendig sind, beginnend mit dem Lesen relevanter Publikationen über die Application Programming Interfaces (API) von PyMed und metapub, bis hin zur Analyse durch ein LLM und Speicherung der Daten in einer lokal gehosteten PostgreSQL-DB. Zusätzlich werden die verwendeten Technologien und Werkzeuge vorgestellt, darunter die Entwicklungs-Umgebung, verschiedene Bibliotheken und Modelle wie LLaMa3, GPT-4 und BERT, sowie Tools wie Docker und Jupyter-Notebook, die für die Datenanalyse, -verarbeitung und -speicherung eingesetzt werden.

2.1 Implementierungs-Prozess

Zur Realisierung des Projekts wurden die notwendigen Schritte geplant und visualisiert als Orientierung für die Entwicklung. Die folgende Abbildung 1 zeigt den Prozess und die Schritte des Projekts und Implementierung, die wie folgt ablaufen:

1. Zuerst werden mit PyMed und metapub relevante Publikationen bzw. Papers gelesen. Relevant sind die Papers, die Resultate haben und Medikationsgruppen getestet haben. Diese Papers sind auf einer DB in PubMed gespeichert, welche von NCBI frei zur Verfügung gestellt wird
2. Die Daten werden lokal vorbereitet, in dem sie angereichert, strukturiert und gesäubert werden
3. Die Daten werden über mehrere Prompts analysiert mithilfe von LLaMa3
4. Diese sendet Medical Records zurück in Form von JSON-Objekten (JavaScript Object Notation)
5. Diese Daten werden abschließend in einer PostgreSQL DB persistiert, die lokal auf einem Docker Container läuft

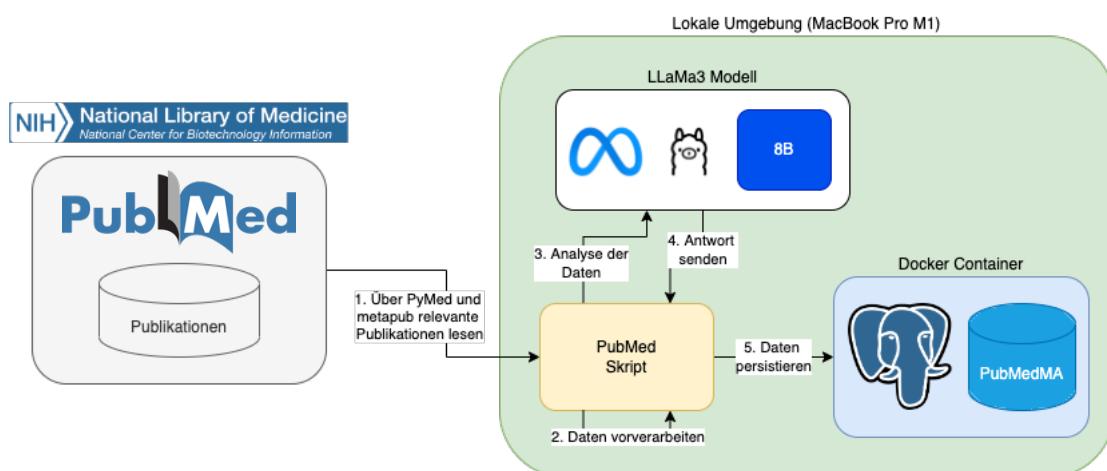


Abbildung 2-1: Visualisierter High-Level Prozess zur Umsetzung des Projektes

2.2 Umgebung, Technologien und Werkzeuge

Zur Realisierung des Projekts werden verschiedene Werkzeuge und Technologien genutzt, die teilweise schon erwähnt wurden. Hier werden sie kurz beschrieben.

Rechner-Umgebung: Als Umgebung zur Entwicklung und Ausführung wird ein MacBook Pro (2021) mit einem M1 Pro Chip sowie 16 GB Arbeitsspeicher verwendet. Bis auf die PubMed-DB wird alles lokal gehostet auf einem privaten Rechner. Auch der gesamte Prozess, bis auf die Datensammlung, erfolgt lokal.

PubMed: Eine von NCBI gehostete DB, die medizinischen Papers hostet und frei zur Verfügung stellt. PubMed wird als Datenquelle für die Analyse benötigt.

API: Mit den Bibliotheken PyMed und metapub, später mehr dazu, werden die API von PubMed genutzt, um die Paper zu lesen und sie für die Analyse zu nutzen.

LLaMa3: Zur Analyse der Paper wird ein LLM verwendet. Dabei handelt es sich um das kleinste LLaMa3 Modell von Meta mit 8 Milliarden Parameter. LLaMa sind Open-Source Modelle, die kostenlos lokal gehostet und genutzt werden können (auch für kommerzielle Zwecke mit großzügigen Rahmenbedingungen).

Python: Als Programmiersprache wird Python 3.11 verwendet. Python bietet zahlreiche Funktionen und Libraries, die für die Entwicklung des Projekts von Vorteil sind.

pandas: Für die Datenmanipulation und -analyse wird die Python-Bibliothek pandas verwendet. pandas bietet leistungsfähige Datenstrukturen wie DataFrames (DF) und eine Vielzahl von Funktionen zur Datenbereinigung, -verarbeitung und -analyse, die für dieses Projekt essenziell sind.

Jupyter-Notebook: Für die interaktive Entwicklung und Ausführung von Code sowie zur Dokumentation und Visualisierung der Ergebnisse wird Jupyter-Notebook eingesetzt. Es ermöglicht die nahtlose Integration von Code, Text, Visualisierungen und Ausgaben in einem einzigen Dokument.

PostgreSQL: Als Datenbankmanagementsystem wird PostgreSQL verwendet. PostgreSQL ist eine leistungsfähige, objekt-relationale DB. Es wird zur Speicherung und Verwaltung der extrahierten und generierten Medical Records genutzt.

Docker: Zur Containerisierung der DB wird Docker verwendet. Docker ermöglicht es, Anwendungen und deren Abhängigkeiten in isolierten Containern zu verpacken, was die Portabilität und Reproduzierbarkeit der Entwicklungsumgebungen verbessert.

GPT-4: Für die Generierung von synthetischen Medical Records und Abstracts, die für die Evaluation der Ergebnisse von LLaMa3 benötigt werden, wird GPT-4 von OpenAI verwendet. GPT-4 ist ein proprietäres und fortschrittliches LLM, das hochqualitative Texte und Daten generieren kann.

BERT: Zur Berechnung von Ähnlichkeiten zwischen zwei Texten wird das BERT-Modell (Bidirectional Encoder Representations from Transformers) eingesetzt. BERT ist ein von Google entwickeltes Modell, das sich durch seine Fähigkeit auszeichnet, den Kontext von Wörtern in einem Text vollständig zu verstehen. Es wird verwendet, um die semantische Ähnlichkeit zwischen Texten der generierten Medical Records zu bewerten. Genau wie GPT-4 wird BERT für die Evaluation der Analyse benötigt.

2.3 Data Collection

In diesem Kapitel wird der Prozess der Datensammlung für das Projekt beschrieben. Es wird erläutert, welche Datenquellen verwendet werden, wie auf diese zugegriffen wird und wie die Daten weiterverarbeitet und gespeichert werden. Beginnend mit einer Einführung in die PubMed-DB des NCBI und den verfügbaren APIs, wird detailliert dargestellt, welche Tools und Bibliotheken zur Datenbeschaffung genutzt werden und wie die Daten schließlich in ein strukturiertes Format gebracht werden. Ziel ist es, eine solide Grundlage für die anschließende Datenanalyse zu schaffen, indem qualitativ hochwertige und umfassende medizinische Daten gesammelt und aufbereitet werden.

2.4 Die Datenquelle

Als Datenquelle für die Analyse und Generierung von medizinischen Records wird die DB PubMed des NCBI verwendet. Das NCBI fördert Wissenschaft und Gesundheit, indem es Zugang zu biomedizinischen und genomischen Informationen bereitstellt. Das NCBI ist Teil der United States National Library of Medicine, einem Zweig der National Institutes of Health. Es wird von der Regierung der Vereinigten Staaten genehmigt und finanziert.

NCBI bietet Online-Informationsressourcen für Biologie und Medizin, darunter die PubMed-DB. PubMed umfasst dabei mehr als 37 Millionen Zitate und Zusammenfassungen biomedizinischer Literatur aus MEDLINE, bio-wissenschaftlichen Fachzeitschriften und Online-Büchern, die frei zur Verfügung gestellt werden. Alle dort vorhandenen Artikel sind veröffentlicht und viele davon sind auch peer-reviewed. Zur Einfachheit werden diese Daten bzw. Zusammenfassungen und Zitate einheitlich als Paper bezeichnet. PubMed ist die wichtigste und einflussreichste DB für medizinische Veröffentlichungen und ein wichtiges Element für die Forschung.

Diese qualitativ hochwertigen und wissenschaftlichen Inhalte stehen frei zur Verfügung und dienen in diesem Projekt als qualitative, fundamentale Grundlage für die Generierung von hochwertigen Records, die auf echten Studien, wahren Daten und Erkenntnissen basieren.

2.5 API zu PubMed und NCBI

NCBI bietet selbst mehrere API, um pragmatisch die Inhalte zu lesen. Jedoch sind diese etwas unhandlich und die Logik zum Anfragen und Erhalten der Inhalte muss selbst programmiert werden und ist damit mit Risiken verbunden. Zusätzlich ist diese nicht gut dokumentiert und eine performante automatisierte Abfrage ist zu kompliziert.

Zum Einfachen Lesen der Papers auf PubMed gibt es mehrere Libraries und Wrappers, die die API von NCBI nutzen. Dies bietet den Vorteil, dass es schon Handler für die Sammlung der Daten gibt und diese nicht selbst programmiert werden müssen. Drei solcher Libraries sind PyMed, metapub und Entrez von Biopython (einer bekannten Sammlung von Tools und Libraries für Aufgaben im Bio- und Chemie-Bereich). Im Folgenden werden sie kurz vorgestellt und analysiert, um abschließend die Library auszuwählen, die in diesem Projekt verwendet wird.

PyMed: PyMed ist eine Python-Bibliothek, die eine einfache Schnittstelle zur PubMed API bietet. Sie ermöglicht das Suchen und Abrufen von PubMed-Datensätzen.

Vorteile:

- Einfache und intuitive Nutzung
- Bündelt Batch-Abfragen und bereitet sowie parst die erhaltenen Daten vor
- Gute Dokumentation und Beispiele

Nachteile:

- Eingeschränkte Funktionalität im Vergleich zu umfassenderen Bibliotheken
- Weniger flexibel für komplexe Abfragen
- Liefert nicht immer alle Daten vollständig zurück

metapub: metapub ist eine umfassende Python-Bibliothek, die verschiedene NCBI-API abdeckt, einschließlich PubMed. Sie bietet fortgeschrittene Funktionen für das Abrufen und Verarbeiten von wissenschaftlichen Publikationen. Sie stellt über eutils geholte Python-Objekte bereit, die PubMed Paper darstellen. Zusätzlich werden weitere Abstraktionsschichten zu weiteren medizinischen DB.

Vorteile:

- Umfassende Funktionalität, deckt mehrere NCBI-APIs ab
- Flexibilität bei der Durchführung komplexer Abfragen
- Unterstützung für die Arbeit mit verschiedenen Datenquellen
- Gutes Handling von PubMed Paper als Python-Objekte von sich aus

Nachteile:

- Komplexer in der Nutzung und Konfiguration
- Steilere Lernkurve im Vergleich zu einfacheren Bibliotheken

Bio.Entrez: Bio.Entrez ist ein Modul innerhalb des Biopython-Pakets, das eine Schnittstelle zur NCBI-DB bietet, einschließlich PubMed. Es ist Teil einer größeren Suite von Bioinformatik-Tools.

Vorteile:

- Teil des umfassenden Biopython-Pakets, das viele Bioinformatik-Tools bietet
- Hohe Flexibilität und Anpassungsfähigkeit
- Unterstützt eine breite Palette von NCBI-DB
- Bietet viele Tools für komplexere Analysen und Anwendungsfälle an, die über dieses Projekt hinausgehen
- Automatisches Error Handling beim Überschreiten von API-Ratenlimits

Nachteile:

- Einarbeitung in Biopython erforderlich
- Komplexer für einfache Abfragen
- Abhängigkeit von einem größeren Paket, was für spezifische Anwendungen überdimensioniert sein kann

Die Wahl der richtigen Bibliothek hängt stark von den spezifischen Anforderungen und dem Erfahrungsniveau des Benutzers ab. Für einfache und schnelle Abfragen ist PyMed am besten geeignet aufgrund seiner Benutzerfreundlichkeit und einfachen Handhabung.

Für umfassendere und komplexe Abfragen bietet metapub eine gute Balance zwischen Funktionalität und Benutzerfreundlichkeit für Nutzer. Für Anwender, die eine breite Palette von Bioinformatik-Tools benötigen, ist Bio.Entrez die beste Wahl aufgrund der umfangreichen Funktionalität und Integration in ein größeres Ökosystem von Bioinformatik-Tools.

Die finale Wahl fällt auf metapub, da es sich gut für die meisten Anwendungen im Bereich der wissenschaftlichen Recherche eignet und ein gutes Fundament für zukünftige Weiterentwicklungen bietet. Außerdem liefert es im Vergleich zu PyMed mehr Daten im Abstract durch die vollständigen Resultate zurück. Bei einem Fall hat PyMed für Results nur einen Teil des Abstracts und Results geliefert, während metapub alles geliefert hat (siehe Anhang A.2).

Kombiniert wird dies mit PyMed, da metapub keine Möglichkeit bietet für eine Query mehrere Papers mit dem kompletten Datensatz zu erhalten. Auch werden die Autoren nicht vollständig ausgeschrieben und das Datum der Veröffentlichung wird nicht genau wiedergegeben (nur das Jahr wird geliefert).

2.6 Vorstellung der Daten

Die Inhalte in PubMed werden strukturiert gespeichert und haben verschiedene Attribute. Die für dieses Projekt relevanten Attribute werden in Tabelle 2-1 vorgestellt:

Tabelle 2-1: Die Daten und Attribute eines Papers in PubMed

Attribut	Beschreibung
PubMed ID	Ein eindeutiger Identifikator (ID) für jedes Paper in der PubMed-DB. Diese ID wird verwendet, um ein spezifisches Paper eindeutig zu identifizieren und zu referenzieren.
Titel des Paper	Der Titel des Peer-reviewed und veröffentlichten Papers. Dieser Titel gibt eine kurze und prägnante Beschreibung des Hauptthemas oder der Forschung des Papers.
Autoren	Die Liste der Autoren, die an der Erstellung des Papers beteiligt waren. Diese wird oft im Format "Nachname, Vorname; Nachname, Vorname" dargestellt und gibt die Namen der beteiligten Forscher an.
Publikationsdatum	Das Datum, an dem das Paper veröffentlicht wurde. Dieses Datum gibt an, wann die Forschung offiziell in einem Journal veröffentlicht und der wissenschaftlichen Gemeinschaft zugänglich gemacht wurde.
Abstract	Eine kurze Zusammenfassung des Papers, die die wichtigsten Punkte, Methoden, Ergebnisse (results) und Schlussfolgerungen (conclusions) der Forschung enthält. Das Abstract bietet einen schnellen Überblick über den Inhalt des Papers.
Results	Eine detaillierte Darstellung der Forschungsergebnisse des Papers. Dieser Abschnitt enthält die Daten, Beobachtungen und Analysen, die aus der Forschung hervorgegangen sind. Kann leer sein.
Conclusion	Die Schlussfolgerungen, die aus den Forschungsergebnissen gezogen wurden. Dieser Abschnitt diskutiert die Bedeutung der Ergebnisse, ihre Implikationen und mögliche Anwendungen oder weiterführende Forschung. Kann leer sein.
Journal	Der Name des wissenschaftlichen Journals, in dem das Paper veröffentlicht wurde. Das Journal gibt den Kontext und die Relevanz des Papers innerhalb eines bestimmten wissenschaftlichen Feldes an.

Im nächsten Schritt werden genau diese Daten gesammelt, verarbeitet und persistiert, bevor sie analysiert werden. Nicht alle Attribute sind immer befüllt, so können Results und Conclusions leer (null) sein.

2.7 Die Sammlung der Daten

Mithilfe der PyMd und metapub Library kann die PubMed DB einfach angefragt werden, um viele Daten zu erhalten. Jedoch wird eine Kombination beider gebraucht, um alle notwendigen Daten zu erhalten. Bis auf Abstract, Results und Conclusions werden alle Daten über PyMed beschaffen.

Hierfür wird über PyMed eine Query an die PubMed-API gesendet, um dazu passende Paper zu filtern und zu erhalten. Die maximale Anzahl an damit erhaltbaren Paper wird ebenfalls gesetzt. In dieser Liste an Paper gibt es für jeden Eintrag eine PubMed ID. Mit dieser wird über metapub eine Anfrage gesendet, um das dazu passende Paper zu erhalten. Mit dieser Methode werden so iterativ die Abstracts gesammelt. In diesen Abstract befinden sich zusätzlich Results und Conclusions als Textbausteine des Abstracts. Diese zwei werden über Regular Expressions (Regex) aus dem Text extrahiert.

In PyMed sind dies zwei eigenständige Attribute, die in einer Antwort als eigene Teile zurückgegeben werden. Da es jedoch ein Problem mit der Vollständigkeit der Daten gibt wird das Rauslesen und Speichern der Attribute Results und Conclusions über metapub und eine Extraktion über Regex durchgeführt.

Alle gesammelten Daten durch die zwei Libraries werden dann in einem DF, dem Master-DF, kombiniert und vollständig gesammelt für Analysen und weitere Vorbereitungen. Dabei werden Duplikate, bestimmt durch die PubMed ID, entfernt.

Im nächsten Schritt wird mit Hilfe von LLaMa3 analysiert, ob medizinische Tests und Gruppen sowie Ergebnisse in den Papers behandelt werden. Falls keine Tests erwähnt werden, werden diese Dokumente mit einer „False“-Flagge markiert. Ein spezieller Prompt, der später genauer beschrieben wird, dient als Grundlage für diese Analyse.

Zusammengefasst läuft der Prozess (visualisiert in Abbildung 2-2) zum Sammeln der Daten wie folgt ab:

1. Erst werden über eine Query über PyMed die "Rohdaten" angefragt inklusive der PubMed ID
2. Diese werden gefiltert: Papers ohne Resultate werden verworfen
3. Mit metapub werden dann über die ID der übrigen Paper iterativ die Abstracts angefragt
4. Über eine Extraktionsfunktion und Regex werden aus dem Abstract Results und Conclusions gelesen
5. Alle Daten werden miteinander gemerged und in einem Master-DF gesichert
6. Mit LLaMa3 wird analysiert, ob medizinische Tests im Paper behandelt werden. Falls nein werden diese mit einer False Flagge markiert

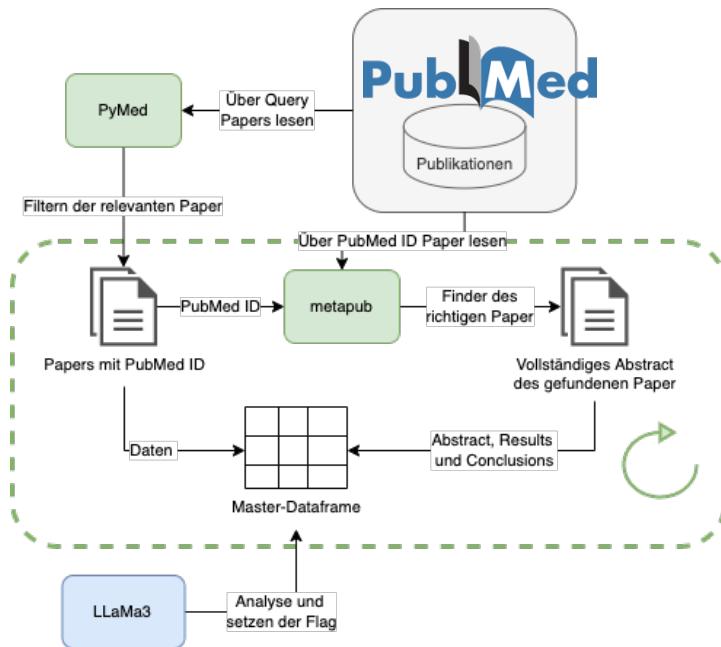


Abbildung 2-2: Prozess der Datensammlung

Für die Erstellung des Master-Dataframes wurde die Krankheit Migräne ausgewählt. Über die Website <https://www.webmd.com/migraines-headaches/migraine-treatments> konnten viele verschiedene Medikationen gefunden werden, um Migräne zu behandeln.

Die Namen der Medikamente wurden als Query verwendet. Zusätzlich werden Migräne und andere Krankheiten gesucht und passende Paper geholt:

- **Migraine:** 2500 Papers angefragt
- **Bronchitis:** 500 Papers
- **Influenza:** 500 Papers
- **Hepatitis:** 500 Papers
- **Sleep Disorders:** 500 Papers
- **Medikamente:** Jeweils 100 Papers

Folgende Medikamente wurden unter anderem angefragt:

- Erenumab
- Topiramate
- Aspirin
- Ibuprofen
- Enumerab
- Paracetamol
- Propranolol
- Siehe Website...

Jede Query-Suche wurde in einem separaten DF gesammelt. Diese wurde dann mit dem Master-DF konkateniert, wobei doppelte Einträge verworfen werden. So konnten insgesamt **2.548 distinkte Paper** gesammelt werden, wobei **1.102 Paper eine True Flagge** erhielten: In diesen Paper geht es um medizinische Studien und es gibt auch medizinische Tests und Ergebnisse. Für die Analyse werden die **1.102 Paper** verwendet.

Obwohl deutlich mehr Paper angefragt wurden als tatsächlich im DF existieren, lässt sich das leicht erklären. Zum einen können verschiedene Queries die gleichen Paper zurückliefern. So können z. B. zwei verschiedene Medikamente die gleichen Paper wiedergeben, da beide in diesem einen Paper getestet wurden. Da doppelte Paper entfernt werden, werden nicht alle Paper übernommen. Ein weiterer Grund ist die Anzahl der passenden Paper für eine Query. Die Query mit der angeforderten Anzahl von 100 Paper gibt nur die Maximalanzahl an mit 100. Das bedeutet nicht, dass auch 100 Paper geliefert werden. So können z. B. selten getestete Medikamente nur 60 statt 100 Paper zurückgeben. Der Hauptgrund ist jedoch das Rausfiltern der nicht relevanten Paper, die keine Resultate haben. So waren im Schnitt **61 %** der erhaltenen Paper ohne Resultate, weshalb sie verworfen wurden.

2.8 Data Preparation

Nach dem Sammeln der Daten werden diese für die Analyse vorbereitet. In der Datenvorbereitung wird eine neues separates DF aufgebaut. Die Daten aus dem Master-DF werden gefiltert und in das neue DF gepackt. Dabei werden Einträge, in denen keine medizinischen Gruppen getestet werden, entfernt. Das geschieht über passende pandas Funktionen und die durch LLaMa3 gesetzte Flag.

Weiter wird das Publikationsdatums einen neuen Datentyp konvertiert, sodass das Persistieren und Analysieren der Daten die Datumsangaben im richtigen Format erhalten. Ohne diese Konvertierung werden diese falsch kodiert und verwendet.

Es wird zusätzlich ein drittes DF gebildet, mit einem kleinen Teilsatz der Daten. Mit diesem werden die Analyse, Prompts und Ergebnisse von LLaMa3 im kleinen Maße getestet. Vorteil davon ist, dass die Analyse nicht lange dauert und so schnell die ersten Medical Records generiert werden. Diese können dann auf Qualität und Korrektheit überprüft werden. Aufbauend auf diesem Feedback wird der Code iterativ verbessert und weiter getestet.

2.9 Persistierung der Daten

Um PubMed nicht ständig neu anzufragen für die gleichen Daten, da mehrere Läufe der Analyse geplant sind, werden die Daten im Master-DF in einer lokalen JSON-Datei (JavaScript Object Notation) gesichert. Falls das Jupyter-Notebook oder der Rechner ausfallen sollten, können so schnell die Daten wiedergelesen werden, ohne den Prozess der Datensammlung zu wiederholen. Außerdem sollen so auch die Daten lokal ohne Internetverbindung weiterverwendet werden.

2.10 Anmerkungen

Beispiele für die erhaltenen Daten und wie ein DF aussieht, sowie weitere Informationen sind im Anhang A zu finden.

3 Prompt Engineering

Prompt Engineering im Kontext von LLM wie LLaMa3 bezieht sich auf die Kunst und Wissenschaft, präzise und effektive Eingabeaufforderungen (Prompts) zu gestalten, um von den Modellen relevante und genaue Antworten zu erhalten im Kontext von NLP-Aufgaben. In diesem Projekt ist Prompt Engineering entscheidend, da es die Qualität und Relevanz der generierten medizinischen Records maßgeblich beeinflusst. Durch gut gestaltete Prompts kann sichergestellt werden, dass LLaMa3 gute Medical Records aus den wissenschaftlichen Publikationen extrahiert und analysiert. Dies ermöglicht eine präzise und konsistente Datenanalyse, die für die Validität der Ergebnisse und somit der DB für die Meta-Analysen unerlässlich ist.

In diesem Abschnitt werden die unterschiedlichen Prompts für LLaMa3 und ihr Aufbau vorgestellt. Diese bestehen aus dynamischen und statischen Teilen. Beim dynamischen Teil handelt es sich meist um Inhalt aus dem gelesenen Paper: Dem Abstract. Die statischen Teile sind die Anweisungen und Regeln zum Verarbeiten des Inhalts. Da es sich in diesem Projekt um eine komplexe NLP-Aufgabe handelt, werden mehrere Prompts benötigt. Diese werden nacheinander aufgerufen, um mit LLaMa3 zu interagieren für eine bessere Analyse. Beispiele für den Inhalt der Prompts sind im Anhang B.

3.1 Prompt zum Filtern von Paper

Dieser Prompt wurde vorhin bei der Vorbereitung der Daten vorgestellt. Er ist dafür zuständig die Papers nach Anwendbarkeit zu filtern. Hier wird entschieden ob in einem Paper Medication Groups oder medizinische Tests vorhanden und getestet werden sowie Resultate beiliegen. Je nach Ergebnis wird eine Flag gesetzt, die entweder True oder False wird. Der Prompt besteht aus vier Teilen: Drei statischen und einem dynamischen:

Statische Teile:

- **Rollenbeschreibung:** Dieser Abschnitt beschreibt die Rolle des Modells als Assistent zur Analyse von medizinischen Abstracts und Ergebnissen, um zukünftige Meta-Analysen zu unterstützen.
- **Konkrete Anweisung:** Diese Anweisung gibt dem Modell klare Handlungsanweisungen: den folgenden Inhalt zu lesen und zu bestimmen, ob in dem Paper Gruppen, Medikamente oder Therapien getestet wurden und Ergebnisse angegeben sind.
- **Regel:** Eine Regel stellt sicher, dass die Antwort des Modells auf eine einfache Ja-oder-Nein-Antwort beschränkt bleibt, um die Aufgabe zu präzisieren und Missverständnisse zu vermeiden.

Dynamische Teile:

Dieser dynamische Teil umfasst den spezifischen Inhalt des Papers, der in den Prompt eingefügt wird. Dies ermöglicht dem Modell, auf Grundlage der spezifischen Daten des Papers zu arbeiten.

Die Antwort von LLaMa sollte zwar nur ein yes oder no sein, jedoch gibt LLaMa sie in verschiedenen Formen zu-rück (YES, yes., Yes, No., NO., usw.). Eine weitere Funktion soll die Antwort richtig handlen und je nach Output True oder False zurückgeben.

Im Anhang B.1 sind einige Beispiele für Prompts und Outputs zu sehen, die im Durchlauf des großen Datensatzes gemacht wurden.

3.2 Prompt zur Bestimmung der Anzahl an Gruppen

Ein weiterer Prompt ist dafür zuständig die Anzahl der Medication Groups zu bestimmen. Es besteht auch aus vier Teilen, drei statischen und demselben dynamischen Teil:

Statische Teile:

- **Rollenbeschreibung:** Dieser Abschnitt beschreibt die Rolle des Modells als Assistent zur Analyse von medizinischen Abstracts und Ergebnissen, um zukünftige Meta-Analysen zu unterstützen.
- **Konkrete Anweisung:** Diese Anweisung gibt dem Modell eine klare Aufgabe: die Anzahl der getesteten Gruppen für einen medizinischen Test aus den nachfolgenden Daten zu extrahieren.
- **Regel:** Diese Regel stellt sicher, dass LLaMa3 nur die Anzahl der Gruppen als Integer zurückgibt und keine zusätzlichen Texte oder Informationen hinzufügt.

Im Anhang B.2 sind einige Beispiele für Prompts und Outputs zu sehen, die im Durchlauf des großen Datensatzes gemacht wurden.

3.3 Prompt zum Generieren von Medical Records

Der nächste Prompt ist dafür zuständig die Medical Records mit den Ergebnissen für die Medication Groups zu extrahieren und diese als JSON-Liste zurückzugeben, um diese im Post Processing zu verarbeiten. Es besteht aus sieben Teilen, fünf statischen und zwei dynamischen:

Statische Teile:

- **Rollenbeschreibung:** Der Prompt beginnt mit einer Beschreibung der Rolle und Aufgaben des Modells in diesem Kontext. Das Modell wird als Assistent definiert, der Abstracts und Ergebnisse aus Papers über Krankheiten und Medikamente liest, um sie für zukünftige Meta-Analysen vorzubereiten.
- **Konkrete Anweisung:** Der nächste Teil gibt eine konkrete Anweisung, was das Modell tun soll. Es wird angewiesen, die Medical Records aus dem dynamischen Inhalt herauszuziehen. Dabei soll auf die Krankheit, die Medication Group, das Medikament, Medical Outcome und Resultat geachtet werden.
- **Hinweis:** Ein zusätzlicher Hinweis informiert LLaMa3, dass mehrere Records möglich sind und die Antworten kurz sein sollen.
- **Regel:** Eine weitere Regel bestimmt, dass die Records als JSON-Liste zurückgegeben werden sollen und gibt die Struktur der JSON-Dokumente vor. Weitere Regeln legen fest, dass unvollständige Records nicht zurückgegeben werden sollen, die Resultate nur den Wert mit der Einheit der Messung enthalten dürfen und dass bei fehlender Angabe der getesteten Gruppe 'No group specified' zurückgegeben wird. Sollten keine Records gefunden werden, soll eine leere Liste zurückgegeben werden. Nur die JSON-Inhalte sollen zurückgegeben werden, keinen weiteren Inhalte sollen generiert werden.

- **Beispiele:** Der letzte Teil enthält Beispiele dafür, wie ein Record aussehen kann. Diese Beispiele dienen als Vorlage, um bessere Ergebnisse zu erzielen.

Dynamische Teile:

- **Anzahl der Medication Groups:** Ein dynamischer Teil ist die Anzahl der gefundenen Medication Groups bzw. Tests für das Paper, die durch den vorherigen Prompt ermittelt wurden.
- **Inhalt des Papers:** Der zweite dynamische Teil ist der Inhalt des Papers selbst, der in den Prompt eingefügt wird.

Im Anhang B.3 sind einige Beispiele für Prompts und Outputs zu sehen, die im Durchlauf des großen Datensatzes gemacht wurden.

3.4 Prompt zur Evaluation von Medical Records

Der letzte Prompt ist dafür zuständig die Medical Records einzeln zu evaluieren. Dabei werden die Outcomes einer Kategorie zugewiesen und die Auswirkung evaluiert.

Statische Teile:

- **Rollenbeschreibung:** Dieser Abschnitt beschreibt die Rolle des Modells als Assistent zur Analyse von medizinischen Abstracts und Ergebnissen, um zukünftige Meta-Analysen zu unterstützen, und die spezifische Aufgabe, die Kategorie und Bewertung des Ergebnisses zu bestimmen.
- **Konkrete Anweisung:** Der nächste Teil gibt eine konkrete Anweisung, was das Modell tun soll. Es wird angewiesen, die Outcome Kategorie und Evaluation zu bestimmen, basierend auf den gegebenen, dynamischen Informationen.
- **Hinweis:** Der Hinweis sagt, dass ein einzelnes JSON-Dokument benötigt wird.
- **Regel:** Regeln geben dem Modell klare Grenzen, welche Informationen in einem JSON-Dokument zurückgegeben werden sollen und wie es aufgebaut ist.
- **Erlaubte Kategorien für Medical Outcome:** Dieser Abschnitt legt fest, welche Kategorien erlaubt sind und dass keine anderen Kategorien erstellt werden dürfen. Jede Kategorie enthält ein Beispiel.
- **Erlaubte Kategorien für Evaluation:** Dieser Abschnitt legt fest, welche Kategorien erlaubt sind und dass keine anderen Kategorien erstellt werden dürfen.

Dynamische Teile:

- **Inhalt des Papers:** Dieser dynamische Teil umfasst den spezifischen Inhalt des Papers, der in den Prompt eingefügt wird. Dies ermöglicht dem Modell, auf Grundlage der spezifischen Daten des Papers zu arbeiten.
- **Record:** Dieser dynamische Teil enthält die spezifischen Informationen des zu analysierenden Records.

Im Anhang B.4 sind einige Beispiele für Prompts und Outputs zu sehen, die im Durchlauf des großen Datensatzes gemacht wurden.

4 Analyse-Prozess

Zur Analyse der Paper wird wie bereits erwähnt ein LLM verwendet. Dabei handelt es sich um LLaMa3. Anfangs wurde auch LLaMa2 mit 7 Milliarden Parameter getestet, jedoch waren die Outputs und Ergebnisse deutlich schlechter. Auch ist LLaMa2 im Vergleich zu LLaMa3 deutlich langsamer. So werden statt **30 Sekunden bei LLaMa3** für ein Record **3 Minuten bei LLaMa2** benötigt. Dies bedeutet eine sechs-mal höhere Ausführungszeit.

Die großen Varianten der beiden Generationen mit 13 und 70 Milliarden Parameter wurden auch getestet, jedoch haben die Anfragen viel zu lange gedauert und die Antworten waren verbuggt. Die lokale Rechenleistung hat bei Weiten nicht ausgereicht.

Das Modell läuft wie eine Art Server und kann über eine API mit Body angefragt werden. Als Antwort werden viele Token zurückgegeben, da ein LLM nach und nach Token generiert und keine ganzen Sätze. Dieser Stream an Tokens muss iterativ verarbeitet werden, um daraus eine lesbare Antwort zu machen. Eine Implementierung im Projekt dient zum Senden von Anfragen an das lokale LLaMa3 Modell und verarbeiten der Response. Die Response wird als JSON zurückgegeben. Weiter wird die Zeit zur Verarbeitung geloggt, um zu sehen, wie lange es dauert, bis LLaMa3 eine vollständige Antwort gibt. So konnten die Ausführungszeiten gemessen werden.

Die Logik wurde so implementiert, dass sie dynamisch funktioniert und so für verschiedene Prompts jeglicher Art wiederverwendet werden kann.

Weiter werden für die Analyse weitere Hilfsfunktionen benötigt, die JSON-Inhalte aus den LLM-Antworten extrahieren. Grund dafür ist, dass das LLaMa3 meist weiteren Text in der Antwort mitgibt, was für die Analyse nicht gebraucht wird und kein JSON hat. Bei der Konvertierung zu JSON bzw. beim Analysieren werden somit Fehler geworfen, da die Response kein JSON ist. Die JSON-Dokumente sind notwendig für das Post Processing und werden in zwei Prompts verwendet. Über die Hilfsfunktionen werden die JSON-Dokumente über Regex identifiziert und erfolgreich extrahiert, sodass sie in Python als Dictionaries oder Listen mit Dictionaries verwendet werden können.

Die weiteren Inhalte neben den JSON-Dokumenten sind nicht erwünscht und werden auch in den Prompts explizit verboten. Trotzdem gibt LLaMa3 in den letzten zwei Prompts meistens weiteren Text mit. Die ersten zwei Prompts funktionieren sehr gut und haben kaum Abweichungen. Ab und zu gibt der erste Prompt zum Setzen der Flag verschiedene Variationen, jedoch hält sich in Grenzen und kann gut abgefangen werden.

Der zusätzlich generierte Text bei den zwei letzten Prompts besteht aus einem einleitenden Satz, der sowas wie ‚*Here is the extracted List of Records as a JSON-List*‘ oder ‚*Here is the extracted JSON-Document*‘. Nach dem JSON-Inhalt folgt eine Begründung für die generierten Inhalte oder warum kein Record gefunden werden konnte (diese Fälle gab es auch selten). Besonders in der Analyse zur Evaluation eines Medical Outcome argumentiert LLaMa3 häufig, wie sich für diese Kategorie und Evaluation entschieden wurde.

Dies kann besonders hilfreich für das Nachvollziehen des Verhaltens sein. Mit dieser Argumentation seitens LLaMa3 als Feedback kann der Prompt weiter verbessert werden, um bessere Ergebnisse zu erhalten. Dies wurde ausgenutzt. Die generierten

Prompts und Antworten, sowie die eigentlichen Daten, werden in einer Datei geloggt. So kann manuell überprüft werden, welche Ergebnisse und Argumente aus der Analyse kommen. Weiter hilft diese Datei dabei Fehler und den Algorithmus zu debuggen.

Die Ausführung dieses Algorithmus ist das Herzstück des Projektes. Dieser sorgt dafür das nacheinander die Prompts ausgeführt werden, um die Paper durch LLaMa3 zu analysieren und Medical Records zu generieren. Die Resultate werden im Post Processing in die DB geschrieben werden. Der Prozess läuft wie folgt ab:

1. Der Algorithmus iteriert über die Einträge im Input-DF, dem vorbereiteten DF. Jeder Eintrag ist dabei ein Paper.
2. Für jedes Paper wird zuerst mit dem Gruppen-Prompt die Anzahl der Gruppen bzw. Tests im Paper bestimmt und zurückgegeben. Als Inhalt bekommt LLama3 den Abstract.
3. Der Abstract und die Anzahl der Gruppe werden zusammen mit dem Records-Prompt wieder an LLaMa3 gesendet. Basierend darauf werden die Medical Records als JSON-Liste gesendet.
4. Die Medical Records werden aus der Antwort extrahiert mit den Hilfsfunktionen.
5. Nun wird über die JSON-Liste und jeden Medical Record iteriert, falls keine leere Liste zurückgegeben wird.
6. Bei einer leeren Liste wird zu Schritt 2 gesprungen mit dem nächsten Paper.
7. Jeder Record wird einzeln und zusammen mit den Resultaten des Papers wieder an LLaMa3 mit dem Evaluations-Prompt gesendet, um den Medical Outcome zu generieren. Basierend darauf wird ein JSON-Objekt geliefert.
8. Das JSON-Objekt wird über eine Hilfsfunktion aus der Antwort extrahiert.
9. Der Record wird mit der Evaluation zu einem Objekt gemerged.
10. Der Record und die Evaluation werden mit weiteren Daten des Papers in der DB über die Post Processing Logikpersistiert.
11. Die Prompts, Antworten, Daten des Papers und die Records samt Evaluation werden in eine JSON-Datei geschrieben.
12. Nach der Records-Schleife ist ein Paper abgearbeitet.
13. Die Schritte 2 bis 112 werden für jedes Paper wiederholt.

Fehlerhafte Durchläufe während der Analyse werden über die pubmed_id in einer Liste gesammelt, um sie für spätere Analysen, Wiederholungen und Verbesserungen zu nutzen. Alle Prompts, Daten und Antworten von LLaMa3 werden in der JSON-Datei gespeichert.

Bei der Analyse wird zusätzlich zwischen zwei Fällen unterschieden, gesteuert durch einen Boolean Input-Parameter. Handelt es sich um die Analyse für die Evaluation, Boolean ist gleich True, wird eine andere Logik zum Speichern der Records in der DB verwendet. Diese sorgt dafür, dass die Records in einer separaten Tabelle persistiert werden. Dazu später mehr.

Nach der Analyse des großen Datensatz im vorbereiteten DF konnten aus 1.035 Paper von 1.102 Medical Records generiert werden. Damit liegt die Quote bei 93,83 %. Die übrigen 67 Paper haben leere Listen zurückgeliefert. Diese Zahl hat je nach Komplett-Durchlauf geschwankt zwischen knapp 1.000 und knapp 1.060 Paper (90,82 % bis 96,12 %). In diesem Fall hat der aller erste Prompt bei manchen Paper die Flag falsch

gesetzt. Hier stellt sich auch die Frage, ob die Paper mit der Flag False auch falsch gesetzt wurden. Davon ist stark auszugehen.

Ein Durchlauf dauert knapp 11 Stunden. Dazu werden noch knapp eine Stunde addiert, um die Flag zu setzen und die Daten vorzubereiten. Fehlerhafte Durchläufe gab es keine, wie in der Abbildung 4-1 zu sehen. Dieser Run war der finale Run und hat die derzeitigen Daten in der DB generiert. Gedauert hat dieser 10 Stunden 40 Minuten.

Hier muss noch angemerkt werden, dass 1.089 Paper analysiert wurden. Die fehlenden 13 Paper wurde für den kleinen Datensatz zum Testen und finalisieren der Pipeline verwendet und wurden dementsprechend weg gelassen, da die Records schon existiert haben.

Das Fehlen von fehlerhaften Läufen in der Liste zeigt, dass die Pipeline stabil läuft und keine bis sehr wenige Fehler wirft. Maximal gab es fünf Fehler bei einem Paper, da ab dem sechsten diese Iteration beendet wird und das Paper und die bis dahin generierten Ergebnisse verworfen werden. Die Prompts und Antworten werden jedoch für das Debugging in die JSON-Datei geschrieben. Die PubMed ID wird in die Liste gepackt für einen zukünftigen verbesserten Durchlauf.

```
[REDACTED] | 1089/1089 [100%] in 10:38:49.0 (0.03/  
1089 runs finished. Data saved to analysis_runs/2024_07_23-00_58_39_analysis_run.json  
2024-07-23 11:37:29.180987 : Analysis done  
Failed runs:  
[]
```

Abbildung 4-1: Analysierte Paper, Dauer und fehlgeschlagene Läufe bei einem Komplett durchlauf des Algorithmus auf den gesamten Datensatz

4.1 Automatisierung durch eine Pipeline

Wie hier zu sehen, handelt es sich um eine kleine Pipeline, die LLaMa3 dazu nutzt, ein Paper vollständig zu analysieren. Durch die implementierte Logik werden die Prompts dynamisch aufgebaut und genutzt, wodurch immer der passende Kontext für LLaMa3 mitgegeben wird, um die bestmöglichen Ergebnisse zu erzielen. Diese vollständigen Prompts ändern sich ständig und ermöglichen es somit, automatisiert durch alle Papers zu iterieren und die Prompts entsprechend anzupassen.

Da die Prompts sequenziell ausgeführt werden, bildet sich eine Pipeline, die ein Paper vom ersten Prompt bis hin zur Persistierung der Daten in der DB vollständig analysiert und verarbeitet, ohne manuelle Zwischenschritte. Alle 13 vorher beschriebenen Schritte werden komplett automatisch durchgeführt, nur der Start benötigt manuellen Input.

Zum Testen der Logik und Arbeitsweise von LLaMa3 und dem Algorithmus wird ein kleiner Satz an Daten genommen. Dieser DF wurde schon im Data Collection Kapitel erwähnt. Damit sollen schnell Analysen gemacht werden, um Ergebnisse und Probleme der Pipeline schnell zu sehen, zu identifizieren und zu verbessern. Hierbei ist aufgefallen, dass unterschiedliche Runs mit denselben Prompts zu unterschiedlichen Ergebnissen führten, die jedoch gleich sind. Nur ihre Repräsentation hat sich geändert, in dem der Satzaufbau umgestellt wurde, Abkürzungen ausgeschrieben wurden und zusätzliche Wörter dazukamen. Ein Beispiel hierfür sind die Medical Outcomes „reduction in MMD“ und „reduction in monthly migraine days (MMD)“. Diese Ergebnisse sind zwar identisch für den Leser, sie unterscheiden sich aber.

4.2 Auswirkungen von Prompt Änderungen

Änderungen in den statischen Teilen der Prompts hatten viel Auswirkungen auf die generierten Medical Records und Prompts. Lange Prompts hatten z. B. negative Auswirkungen auf die Verarbeitungszeit und Qualität der Analyse. Dies war in der Kombination der zwei Prompts für das Extrahieren von Medical Records und der Bewertung zu beobachten. Durch die iterative Analyse der Records innerhalb der schon großen Schleife wurde zwar die Analysezeit massiv erhöht, jedoch waren die Ergebnisse hochwertiger, da mehr und besserer Fokus bzw. Aufmerksamkeit auf die wichtigen Daten gesetzt wurde und es weniger Ablenkung gab durch viele andere Token.

Regelbasierte Prompts führten zu besseren und qualitativ hochwertigeren Ergebnissen. Durch das Festlegen spezifischer Strukturen, wie zum Beispiel das Format von JSON-Objekten, konnte die Wahrscheinlichkeit erhöht werden, dass die generierten Outputs konsistent und fehlerfrei waren. Diese Strukturvorgaben halfen, die erwarteten Formate beizubehalten und damit das Post Processing der Daten zu erleichtern.

Konkrete Anweisungen innerhalb der Prompts erwiesen sich als äußerst hilfreich, um den Fokus auf die wesentlichen Aufgaben zu lenken und Ablenkungen zu minimieren. Klare und präzise Anweisungen ermöglichten es LLaMa3 effizienter und zielgerichtet zu arbeiten, was die Qualität der Ergebnisse deutlich verbesserte. Dies wurde besonders bei der Extraktion relevanter Informationen aus viel Text deutlich.

Die in den Prompts enthalten Beispiele spielten eine entscheidende Rolle bei der Verbesserung der Datenverarbeitung. Durch die Bereitstellung von Beispielen konnte das System besser verstehen, wie mit den Daten umzugehen ist. Dies führte zu einer besseren Strukturierung und einer höheren Qualität der generierten Ergebnisse. Die Beispiele dienten als Vorlage und Referenz, was zu konsistenteren und verlässlicheren Outputs führte.

Um schnelles Feedback und erste Ergebnisse zu erhalten, wurde ein kleiner Datensatz verwendet. Dies ermöglichte eine zügige Analyse und lieferte schnell verwertbare Erkenntnisse. Durch dieses Vorgehen konnte die Effizienz der Entwicklungsschleife erhöht werden, da Probleme und Optimierungspotenziale schneller identifiziert und umgesetzt wurden. Die Prompts konnten iterativ weiterentwickelt werden.

Dieser iterative Ansatz war essenziell, um die Qualität der generierten Medical Records kontinuierlich zu steigern. Durch schrittweise Anpassungen und Optimierungen konnten Fehler identifiziert und behoben werden. Dieser iterative Prozess stellte sicher, dass die Prompts immer besser auf die spezifischen Anforderungen und Herausforderungen der Analyse abgestimmt waren.

Zusammenfassend lässt sich sagen, dass die gezielten Änderungen und Optimierungen der Prompts zu einer signifikanten Verbesserung der Analysequalität und Effizienz führten. Regelbasierte Ansätze, konkrete Anweisungen und Beispiele sowie ein iterativer Entwicklungsprozess erwiesen sich als Schlüsselfaktoren für den Erfolg.

4.3 Weitere Anmerkungen

Einige Informationen und Bilder zur Analyse befinden sich im Anhang C.

5 Post Processing

Im Post Processing geht es um die Implementierung der Logik und Umgebung, mit der die Medical Records nach der Analyse verarbeitet und persistiert werden. Wie anfangs erwähnt wird eine lokale PostgreSQL-DB verwendet, die in einem Docker-Container läuft. Diese hat zwei Tabellen, medical_records und medical_records_evaluation, die identisch aufgebaut sind.

Der Aufbau der Tabelle und damit eines Medical Records sieht wie folgt aus:

Attribut	Beschreibung	Datentyp
id	Eindeutige Nummer des Datenbankeintrags	<i>String</i>
pubmed_id	Eindeutige Nummer des Peer-reviewed und veröffentlichten Papers	<i>String</i>
papername	Name des Peer-reviewed und veröffentlichten Papers	<i>String</i>
authors	Autoren des Papers im Format: Nachname, Vorname;; Nachname, Vorname	<i>String</i>
journal	Name des Journals, in dem das Paper veröffentlicht wurde (optional)	<i>String</i>
publication_date	Datum, an dem das Paper veröffentlicht wurde	<i>Date</i>
disease	Name der Krankheit, um die es in dem Paper handelt	<i>String</i>
medication_group	Die getestete Gruppe, falls vorhanden.	<i>String</i>
therapy_or_medication	Benutzte Therapie bzw. benutztes Medikament für die Behandlung der Krankheit bei einer getesteten Medikationsgruppe	<i>String</i>
medical_outcome	Art des Ergebnisses oder gemessener Endpunkt der Medikation in der getesteten Gruppe (Beobachtung oder Messung der Wirksamkeit bzw. des Einflusses)	<i>String</i>
result	Spezifische Zahlen oder Prozentsätze des Ergebnis der Medikation in der getesteten Gruppe (Quantitative Messung aus der Messung des Medical Outcome)	<i>String</i>
outcome_category	Kategorie des Outcome	<i>String</i>
outcome_evaluation	Angabe, ob das Outcome zu einer Verbesserung oder Verschlechterung führt	<i>String</i>

5.1 Interaktionen mit der Datenbank

Die Verbindung zur DB wird über pyscope implementiert, einer Python-Library für Interaktionen mit PostgreSQL-DB zum Lesen und Schreiben von Daten.

In diesem Projekt gibt es insgesamt drei SQL-Queries (vorausgesetzt das die zwei Tabellen schon existieren). Eine SQL-Query wird zum Persistieren der Medical Records in der Tabelle genutzt. Die dafür benötigte SQL-INSERT-Query nimmt alle Daten auf, weist sie den richtigen Spalten zu und wird über den Cursor der Verbindung ausgeführt und committed. Die Query ist dynamisch und passt sich den Daten und Records an.

Zusätzlich wird über ON CONFLICT dafür gesorgt, dass identische Datensätze (bestimmt über pubmed_id, disease, therapy_or_medication, medical_outcome und result) nicht nochmal gespeichert werden und so die Meta-Analysen verfälschen. Dies ist notwendig, da verschiedene Queries bei der Datensammlung die gleichen Papers zurückliefern und verschiedene Durchläufe so identische Medical Records in der DB speichern. Dies ist jedoch nur eine einfache Möglichkeit doppelte Records zu verhindern, da

hier nur identische Records abgefangen werden. Wie bei der Analyse beschrieben gibt LLaMa3 zwar ähnliche bzw. gleiche Records zurück, jedoch unterscheiden sie sich teilweise in den Repräsentationen.

Die zweite Query läuft analog zur ersten, mit dem Unterschied, dass in eine andere Tabelle geschrieben wird. Dabei handelt es sich um die Tabelle zur Evaluation. Die letzte Query ist eine READ-SQL-Query zum Lesen der Daten aus der Evaluations-Tabelle in ein Dataframe zur Evaluation, was später beschrieben wird.

5.2 Erweiterte Pipeline: Vom LLM zur Datenbank

Die Logik zur Interaktion mit der DB erweitert bzw. finalisiert die Pipeline und versorgt sie mit den notwendigen Funktionen, um mit der DB zu interagieren und die Medical Records zu persistieren. Die Funktionen realisieren den zehnten Schritt des Analyse Prozess und übernehmen damit das Post Processing, da sie die Daten nach der vollständigen Analyse verarbeiten.

5.3 Generierte Records und ersten Abfragen

Das Post-Processing und die INSERT-Queries verliefen ohne Probleme. Die Daten konnten alle in einer Tabellepersistiert werden. Aus den 1.035 Paper konnten insgesamt 3.659 Medical Records bzw. Datensätze generiert werden. Diese Zahl schwankt je nach Lauf zwischen 3.500 und 3.800 Records. Die Schwankungen kommen unter anderem davon, dass LLaMa3 pro Durchlauf hin und wieder leere Listen für Paper zurückgibt, da sie entweder keine haben oder LLaMa3 sie bei diesem Durchlauf nicht entdecken konnte.

Das Lesen der Daten aus der Tabelle über SELECT-Queries haben ebenfalls problemlos funktioniert. Auch komplexere Abfragen, wie GROUP BY und Aggregationen wie COUNT konnten ausgeführt werden. Dazu später in Ergebnisse mehr.

	id [PK] integer	pubmed_id integer	paper text	authors text	journal character varying (255)	publication_date date	disease character var	medication_group character varying (255)	therapy_or_mi character vary	medical_outcome character varying (255)	result character vary	outcome_category character varying (255)	outcome_evaluation character varying (255)
52	1486	31743114	Comparing Nonopioids Ver...	Sin, Billy; Sikorska, Ga...	American journ...	2019-11-20	Acute pain	2323 patients	Opioids (...	Pain reduction	2 trials rep...	Symptom Intensity...	significantly improved
53	1487	31743114	Comparing Nonopioids Ver...	Sin, Billy; Sikorska, Ga...	American journ...	2019-11-20	Acute pain	2323 patients	Nonopiod...	Pain reduction	1 trial rep...	Efficacy Measures (...	no significant differen...
54	1488	37760827	Potentially Inappropriate U...	Machado-Duque, Man...	Biomedicines	2023-09-28	Migraine	Patients diagnosed with...	Antimigrai...	Frequency of opioid use	20.6% patt...	Medication Usage	significantly worse
55	1489	37760827	Potentially Inappropriate U...	Machado-Duque, Man...	Biomedicines	2023-09-28	Migraine	No group specified	Opioids	New opioid use during ...	22.8% patt...	Medication Usage	significantly worse
56	1490	37460781	Association Between Migr...	Kim, Yonghwan; Shin, ...	Clinical drug inv...	2023-07-18	Migraine	Male patients	Triptans a...	Risk of ischemic CCVDs	1.39	Physiological Measu...	significantly worse
57	1491	37460781	Association Between Migr...	Kim, Yonghwan; Shin, ...	Clinical drug inv...	2023-07-18	Migraine	Female patients	Triptans a...	Risk of ischemic CCVDs	1.53	Physiological Measu...	significantly worse
58	1492	37460781	Association Between Migr...	Kim, Yonghwan; Shin, ...	Clinical drug inv...	2023-07-18	Migraine	No group specified	Triptans a...	Risk of ischemic CCVDs	0.86	Physiological Measu...	significantly worse
59	1493	36242871	Evaluating the correlation ...	Esmaili, Sara; Alizade...	Journal of clin...	2022-10-16	Migraine	Patients with migraine a...	No therap...	Subclinical atheroscler...	64.9%	Physiological Measu...	no significant differen...
60	1494	36242871	Evaluating the correlation ...	Esmaili, Sara; Alizade...	Journal of clin...	2022-10-16	Migraine	Patients with migraine	No therap...	Increased IMT and low...	p = 0.247 ...	Physiological Measu...	no significant differen...
61	166	37940860	One-year prospective real...	Lanter-Minet, M; Fabr...	The journal of h...	2023-11-09	Migraine	No group specified	Erenumab	reversion from chronic...	57.3%	Efficacy Measures	significantly improved
62	167	37940860	One-year prospective real...	Lanter-Minet, M; Fabr...	The journal of h...	2023-11-09	Migraine	No group specified	Erenumab	transition from medica...	46.5%	Efficacy Measures	significantly improved
Total rows: 3659 of 3659 Query complete 00:00:00.195													

Abbildung 5-1: Teilausschnitt der 3.659 Medical Records in der DB-Tabelle

6 Evaluation

Wie vorher erwähnt gibt es zwei verschiedene Funktionen, die das gleiche ausführen, jedoch auf eine andere Tabelle in der DB. Dies ist notwendig zur Evaluierung der Ergebnisse und Berechnung der *Accuracy* bzw. *Genauigkeit* der Daten.

Um sicherzustellen, dass die Antworten eines LLM wie LLaMa3 korrekt und zuverlässig sind, wenn es um medizinische Aufzeichnungen geht, ist ein strukturiertes Evaluationsverfahren mit Quality Checks erforderlich. Dieses Verfahren muss verschiedene Aspekte der Genauigkeit und Konsistenz der Antworten bewerten. Nachfolgend wird das Evaluations-Konzept für dieses Projekt vorgestellt.

6.1 Evaluationskonzept

Definieren der Qualitätskriterien: Qualitätskriterien sollten klar definiert sein und beinhalten in diesem Projekt:

- **Korrektheit der Informationen:** Übereinstimmung der Antworten mit den medizinischen Aufzeichnungen.
- **Vollständigkeit:** Die Antwort soll alle wesentlichen Informationen enthalten.
- **Konsistenz:** Konsistente Antworten bei wiederholten Anfragen oder ähnlichen Fällen.
- **Verständlichkeit:** Die Antwort sollte klar und verständlich formuliert sein.

Automatisierte Überprüfung: Die Evaluation sollte automatisiert und schnell erfolgen, um Fehler bei der manuellen Evaluierung zu minimieren. Hierfür werden Tools benötigt, um die LLM-Antworten zu überprüfen. Ausgeführt wird dies durch reale Records, einen Evaluationsalgorithmus und BERT zum Bewerten der Ähnlichkeit der Ergebnisse mit den echten Daten.

Menschliche Überprüfung: Menschliche manuelle Überprüfung ist entscheidend, um sicherzustellen, dass die Antworten korrekt und leicht verständlich sowie sinnhaftig sind. Ein Mensch bewertet die Antworten hinsichtlich ihrer Genauigkeit, in dem die echten Daten und das Paper gelesen und mit den generierten Daten verglichen werden.

Testfälle und Szenarien: Es werden Testfälle bzw. Records und Abstracts erstellt, die typische und atypische medizinische Fälle abdecken. Damit kann überprüft werden, ob LLaMa3 die Records richtig in ihrem Inhalt und Anzahl erkennt und wiedergibt. Dafür werden zwei Ansätze verwendet:

- **Synthetische Datensätze:** Generierung von synthetischen, aber realistischen Medical Records zur Bewertung.
- **Edge-Cases:** Einbeziehung von seltenen, komplexen und unrealistischen Fällen, um die Robustheit des Modells zu testen.

Metriken: Mit Metriken wird die Leistung des Modells quantifiziert und bewertet. Dafür wird die bereits erwähnte Genauigkeit bzw. Accuracy verwendet. Diese gibt den Prozent-Anteil der richtig getroffenen Antworten wieder, basierend auf den realen Daten.

Mit diesem Konzept soll die Leistung bzw. Performance und Genauigkeit der Antworten des Modells und der Pipeline bewertet werden.

6.2 Umsetzung der Evaluation

Nachfolgend wird der Prozess zur Umsetzung der Evaluation beschrieben. In Abbildung 6-1 wird dieser Prozess visualisiert.

Zur Evaluation der Performance von LLaMa3 wurden Daten zur Evaluation beschafft. Dabei wurden mithilfe von GPT-4 und einem passenden Prompt synthetische Daten und Edge-Cases als Medical Records generiert und in einer Excel-Tabelle gespeichert. Diese Daten werden als reale Daten bzw. **reale Records** bezeichnet. 100 Records werden dabei erstellt, wobei ein Paper wie im echten Fall mehrere Records haben kann.

Mit diesen Daten wurden wieder mit GPT-4 insgesamt 46 Abstracts generiert. Als Beispiel und als Vorlage wurden einige echte Abstracts von PubMed mitgegeben, um die Abstracts so realistisch wie möglich zu machen. LLaMa3 sollte diese genau wie die richtigen Abstracts in einem normalen Durchlauf analysieren und daraus Records generieren. Diese Records werden später mit den realen Daten verglichen.

Aus den Daten der Tabelle (Autoren, Titel, PubMed ID usw.) und den Abstracts wurde ein DF für die Evaluation gebildet. Auf diesen DF folgte der Analyse-Prozess mit den 13 Schritten zum Generieren von Medical Records, die nun als **vorhergesagte Records** bezeichnet werden. Der Boolean Input-Parameter wurde hierfür auf True gesetzt, wodurch die Records in die Tabelle `medical_records_evaluation` geschrieben wurden im Post Processing. Die generierten Records wurden dann über die Lese-Funktion in eine neue Tabelle bzw. in ein analysis-DF gelesen.

Iterativ wurden dann die Ergebnisse mit den echten Daten aus der Excel-Tabelle in ihrer Ähnlichkeit verglichen und evaluiert. Das erfolgte mit mehreren Funktionen und Modellen: BERT, einer Funktion zum Vergleich der Anzahl der Records und einer weiteren zum Vergleich der Anzahl an korrekt geflaggten Paper. Die Resultate, in diesem Fall die Ergebnisse der Vergleiche zwischen den Inhalten der realen und vorhergesagten Records, wurden in einer JSON-Evaluations-Datei persistiert.

Die Ergebnisse der Evaluation können so jederzeit genutzt werden für verschiedene Analysen, Bewertungen und/oder Visualisierungen.

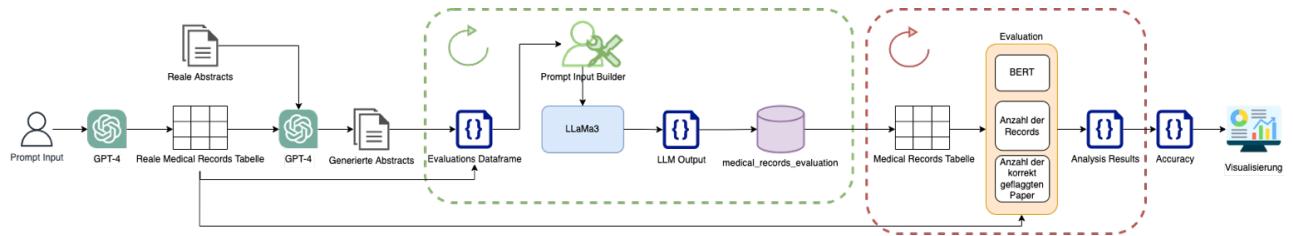


Abbildung 6-1: Prozess zur Evaluation der LLaMa3-Analyse und dessen Antworten

6.3 Evaluations-Algorithmen

Für die Evaluation der Ergebnisse und der Performance des LLM wird wie vorher erwähnt die Genauigkeit bzw. Accuracy verwendet. Diese sagt aus wie oft das LLM von n Versuchen richtig vorhergesagt/klassifiziert hat, also wie oft korrekt vorhergesagt wurde. In diesem Projekt wird lediglich diese Metrik verwendet. Mithilfe von Algorithmen soll die Accuracy für mehrere Evaluationen berechnet werden.

Für die Bestimmung der Accuracy der korrekt geflaggten Paper wird der erste Filter-Prompt und LLaMa3 genutzt. LLaMa3 setzt die Flags auf entweder True oder False. Die Anzahl der Paper mit der True-Flag wird dann durch die Anzahl an Paper geteilt (da in diesem Fall alle Paper medizinische Tests und Gruppen haben). Die Accuracy gibt hier den Teil der korrekt klassifizierten Paper an. Folgende Formel beschreibt die Berechnung:

$$\text{Accuracy} = \frac{\text{Anzahl der Paper mit True Flag}}{\text{Anzahl der Paper}}$$

Die Bestimmung der Genauigkeit der Anzahl der Records ist einfach. Hier wird die Anzahl der vorhergesagten Records mit der Anzahl der realen Records dividiert. Da LLaMa3 eventuell nicht alle Records findet sinkt die Genauigkeit. Diese ist einfach zu berechnen, da in diesem Konzept es genau 100 Records gibt. Folgende Formel beschreibt die Berechnung:

$$\text{Accuracy} = \frac{\text{Anzahl der vorhergesagten Records}}{\text{Anzahl der realen Records}}$$

Zusätzlich wird überprüft, wie hoch der Anteil der Paper ist, bei denen die Anzahl der Records richtig generiert wurde. Dies wird ebenfalls als Accuracy angegeben.

Für die Berechnung der Accuracy der generierten Records und damit der eigentlichen Performance des Modells wird mehr Logik benötigt. Hierfür müssen die generierten Attribute (disease, medication group, therapy or medication, medical outcome, result, outcome category und otcome evaluation) mit den realen Daten auf Ähnlichkeit bewertet werden. Dies erfolgt über BERT. Die Ähnlichkeit wird als Prozentangabe zurückgegeben. Überschreitet diese einen gewissen Threshold, einer Art Schwellenwert, sind die Daten ähnlich bis identisch bzw. gleich. Damit lag LLaMa3 bei der Vorhersage richtig und die Accuracy steigt. Diese Ähnlichkeitsmessung muss für alle Attribute und alle vorhergesagten Records ausgeführt werden, um so alle richtigen Treffer herauszufinden.

Für die Anzahl der richtigen Treffer pro Feld, um damit die Accuracy des Modells zu bestimmen, wird ein Dictionary mit Hilfsvariablen gepflegt (siehe Abbildung 6-2). In diesem wird die Anzahl der Treffer (correct) gespeichert und später für Berechnungen verwendet. Es wird über alle Records in einem Algorithmus iteriert und die Attribute der zusammengehörenden realen und vorhergesagten Records werden miteinander verglichen. Ergibt die Ähnlichkeit True wird die passende Hilfsvariable für das Attribut inkrementiert. Dabei wird ein Vergleich True, wenn der Threshold übertroffen wird. Dies wird für jedes Attribut gemacht. Sollten beim bei einem Record alle Attribute ähnlich sein und True ergeben, sind die Records komplett identisch und LLaMa3 hat es richtig erkannt. Hierbei wird die Hilfsvariable completely_true inkrementiert.

```
{
    "records": 100,
    "completely_correct": 19,
    "disease_correct": 91,
    "medication_group_correct": 56,
    "therapy_or_medication_correct": 97,
    "medical_outcome_correct": 99,
    "result_correct": 72,
    "outcome_category_correct": 71,
    "outcome_evaluation_correct": 85
}
```

Abbildung 6-2: Beispiel Dictionary mit Hilfsvariablen für die Evaluation

Mit dem BERT-Modell wird dieser Algorithmus implementiert. Dieser dient zur Berechnung der Textähnlichkeit mit Hilfe eines Deep Neural Network. BERT ist ein hochmodernes Modell für viele Aufgaben im Bereich NLP. Insgesamt ermöglicht dieser, die Ähnlichkeit zwischen zwei Texten auf eine hochentwickelte Weise zu schätzen, indem die semantischen Merkmale der Texte durch das BERT-Modell extrahiert und verglichen werden. Dies ist besonders nützlich in Anwendungen wie semantischer Textvergleich, Textklassifikation und anderen NLP-Aufgaben.

Der tatsächliche Vergleich zwischen den Records ist etwas komplizierter als oben beschrieben, da die Reihenfolge der Records in der Tabelle nicht unbedingt mit der in Excel übereinstimmt. Die zusammengehörenden Records müssen erst gefunden werden. In diesem Fall muss für ein Paper jedes reale Record mit jedem vorhergesagten Record verglichen werden. Hierfür wird ein Score berechnet: Dieser ist die Summe aller Ähnlichkeitswerte der verglichenen Attribute. Die Ähnlichkeitswerte werden durch BERT bestimmt.

Für die Umsetzung werden zuerst ein Paper und dessen Records ausgewählt. Über diesen und dessen reale Records wird iteriert. Ein realer Record entspricht dem synthetisierten Record in der Excel-Tabelle. Jeder reale Record wird mit jedem vom LLaMa3 vorhergesagtem Record auf Ähnlichkeit verglichen und der Score berechnet. Der Vergleich findet zwischen den zu vorhersagbaren Attributen statt. Der Record mit dem besten Score für einen realen Record wird in einem separaten Dictionary zwischengespeichert inklusive den Ähnlichkeitswerten für jedes Attribut. Alle Läufe werden zusätzlich in einer weiteren Liste gespeichert.

Diese Logik wird für alle realen Records für alle Papers durchgeführt. Am Ende werden die Ergebnisse in einem JSON-Dokument zwischengespeichert. Diese beinhaltet eine Liste mit komplexen JSON-Objekten für jeden realen Record. In so einem Objekt befinden sich neben den Daten des realen Records das beste vorhergesagte Record und die Ähnlichkeitswerte für jedes Attribut sowie der Score. Zusätzlich befindet sich eine Liste mit allen vorhergesagten Records, samt den Ähnlichkeitswerten und Scores, in diesem Objekt. Diese Daten werden in einer JSON-Datei gespeichert.

Die JSON-Datei kann so jederzeit gelesen werden für Analysen und Evaluationen. Dies hat den Vorteil dass die Ergebnisse immer da sind, auch wenn Jupyter-Notebook oder der Computer herunterfahren. Mit der vollständigen Erfassung aller Records und Ähnlichkeitswerte, sowie der Speicherung der Ähnlichkeitswerte als rohe, unbearbeitete Prozentwerte (von 0 bis 1), gibt es für die Evaluation alle fundamentalen Daten. Evaluationen und auch Visualisierungen können so flexibel implementiert werden. Die Genauigkeit ist ebenso flexibel bestimmbar. Mit Thresholds und den Ähnlichkeitswerten kann die Accuracy der Analyse flexibel berechnet werden und sich an die Daten, Anforderungen und Attribute anpassen. Als Nachteil ist hier zu erwähnen, dass diese Logik selbst implementiert werden muss.

Zur Visualisierung der Ergebnisse werden die Ergebnisdaten aus der Datei gelesen. Die Accuracy wird für jedes Attribut als Prozentangabe berechnet, in dem die Anzahl der richtigen Vorhersagen durch die Anzahl der realen Records geteilt wird. Diese Anzahl beträgt 100, wodurch sie leicht selbst zu berechnen ist. Alle Genauigkeiten werden in einem Balkendiagramm dargestellt, um schnell zu erkennen, welche Attribute gut und nicht gut von LLaMa3 erkannt wurden.

6.4 Bewertung der Evaluation

In den folgenden Abschnitten werden die Ergebnisse der Evaluations-Algorithmen vorgestellt und bewertet.

6.4.1 Accuracy der korrekt klassifizierten Paper

Wie in Abbildung 6-3 zu sehen, hat LLaMa3 alle Paper richtig klassifiziert und erkannt, dass in allen synthetischen Papern medizinische Tests durchgeführt wurden. Die Genauigkeit liegt hier bei 100 %.

Je nach Run können auch nur 45 Paper richtig klassifiziert werden. Die Accuracy liegt hier dann bei 97,82 %. Unter 45 Paper ist die Zahl nie gefallen.

```
Number of papers with medication groups: 46
Total count of rows: 46
Accuracy: 100.00%
```

Abbildung 6-3: Ergebnis der Accuracy für die Anzahl der korrekt geflaggten Paper

6.4.2 Accuracy der Medical Records

Auch die Anzahl der generierten Records wird überprüft, um so zu bestimmen, ob LLaMa3 potenziell alle Records erkannt hat. Nur potenziell, da die Inhalte falsch sein können. In den generierten realen Daten wurde die Anzahl der Records bestimmt, wobei die Zahl zwischen einem und sechs Records pro synthetisches Paper schwankt.

Wie in Abbildung 6-4 zu sehen, wurde bei 44 von 46 Papers die korrekte Anzahl an Medical Records erkannt und generiert. Bei den zwei fehlerhaften Papers wurden bei Einem nur drei von sechs Records generiert und beim Zweiten sogar zwei statt ein Record generiert. Somit liegt hier die Accuracy bei 95,65 %.

Von 100 Records hat LLaMa3 98 generiert und damit eine Accuracy von 98 % erreicht, wobei diese Zahl nicht eindeutig ist, da bei einem Paper z. B. mehr Records generiert wurden, als es geben sollte. Damit ist klar, dass mindestens ein Record fehlerhaft ist, da LLaMa3 hier halluziniert hat, oder ein Record extrahiert hat, was im Paper in diesem Kontext nicht existiert. Möglicherweise wurden die Ergebnisse des einen Records in zwei aufgeteilt. Dieser Fall wird später bei der Evaluierung der Accuracy der generierten Inhalte einen negativen Einfluss auf das Ergebnis haben.

Bei diesem Durchlauf handelt es sich um einen vergleichsweisen guten Lauf. In anderen Läufen schwankte die Anzahl der Papers mit der gleichen Anzahl an Records zwischen 39 und 45 Papers (84,78 % und 97,82 %). Öfters wurden 42 Paper richtig extrahiert (91,30 %).

Die Anzahl der Records schwankte zwischen 84 und 100 Records, wobei beide Grenzwerte nur selten bzw. einmal vorgekommen sind. Häufiger kamen 86 bis 94 Records als Ergebnis. Über 100 Records wurden bei keinem Durchlauf generiert.

Trotz Schwankungen hat LLaMa3 hier gute Ergebnisse geliefert und die Anzahl der Records gut erkannt.

```

Not identical for PubMed_ID 50101: Real: 6 Records, Pred: 3 Records
Not identical for PubMed_ID 90909: Real: 1 Records, Pred: 2 Records
Count of papers with matching records: 44
Real count of records: 100
Count of generated records: 98
Total count of papers: 46
Accuracy matching Papers: 95.65%
Accuracy generated records: 98.00%

```

Abbildung 6-4: Ergebnisse der Accuracy für die Evaluation der Anzahl an Records

6.4.3 Genauigkeit der Records

Die Genauigkeit der Inhalte eines Records kann über die Thresholds für die Ähnlichkeit berechnet und geändert werden. Dabei gibt es mehrere Probleme und einen Tradeoff.

Beim **Tradeoff der Ähnlichkeit** bedeutet ein hoher Threshold, dass die Genauigkeit geringer ist, aber die Ergebnisse klarer und eindeutiger sind. Im Gegensatz dazu führt ein niedriger Threshold zu einer höheren Genauigkeit, jedoch auf Kosten weniger eindeutiger Ergebnisse. Dieser Kompromiss muss sorgfältig abgewogen werden, um die Balance zwischen Präzision und Klarheit der Ergebnisse zu finden.

Dies ist jedoch nur ein naiver Ansatz, da damit mehr weitaus mehr Probleme verbunden sind als anfangs vermutet. Identische Inhalte werden mit einer niedrigen Ähnlichkeit bewertet aufgrund von unterschiedlichen Repräsentationen. Manchmal werden Abkürzungen ausgeschrieben, Sätze umgebaut oder verschiedene Wörter/Synonyme verwendet, weshalb die Ähnlichkeit sinkt. Ein Mensch kann hier viel genauer klassifizieren, erfordert jedoch mehr Zeitaufwand. Dagegen können unterschiedliche Inhalte eine hohe Accuracy haben, da sie einige Wörter teilen und nur ein Wort für den Unterschied sorgt. Dies wird gleich näher mit Beispielen betrachtet.

Da sich die Attribute stark in ihren Inhalten und Freiheiten unterscheiden, kann kein einheitliches Threshold gewählt werden. So sollten outcome evaluation und outcome category einen sehr hohen Threshold haben, da ihre möglichen Werte vorgegeben werden. Hier gibt es wenig Freiheit, daher muss die Ähnlichkeit so hoch wie möglich sein, damit die Werte identisch sind.

Disease und medication haben ebenfalls einen relativ hohen Threshold, da hier es weniger Freiheiten gibt. Es gibt sehr wenige bis kaum Synonyme für diese und es reichen meist eins bis zwei Wörter diese zu beschreiben. Da Abkürzungen vorkommen können wird der Threshold nicht zu hoch gesetzt.

Medical outcome, medication group und results haben die höchsten Freiheitsgrade durch verschiedene Repräsentationen. Daher sollte ein etwas niedrigerer Threshold gewählt werden, wobei der Trade-Off nicht vergessen werden darf.

Die besten Thresholds wurden basierend auf den Counter bzw. der Accuracy für jedes Attribut bei verschiedenen Thresholds bestimmt (siehe Tabelle 6-1). Es wurde versucht eine Balance zwischen hoher Accuracy und möglichst hohem Threshold, also hoher Ähnlichkeit, zu erreichen. Jedoch ist dies nicht möglich, da z. B. manche Vergleiche trotz der gleichen Krankheit nur einen Ähnlichkeitswert von 0.27 hatte. Ein Beispiel dafür

ist der Vergleich zwischen „Arthritis“ und „Rheumatoid Arthritis“ (siehe Abbildung 6-5). Für einen Menschen und einen Experten in diesem Bereich sind diese Antworten identisch. Laut BERT jedoch nicht. Das Risiko das tatsächlich falsche Vorhersagen mit dabei sind, wenn das Threshold auf 0.25 gesetzt wird, und damit die Accuracy verfälscht wird, ist hierbei jedoch zu hoch, weshalb dieser Trade-Off eingegangen wird.

Weiter wurde beachtet, dass der Threshold hoch genug ist, um unterschiedliche Texte, die jedoch eine ähnliche Idee verfolgen und gleiche Wörter miteinander teilen, nicht als identisch zu klassifizieren. Ein Beispiel sind die Texte bzw. Klassifizierungen „significantly improved“ und „significantly worse“. Beide teilen sich „significantly“ und haben die gleiche Idee: Etwas hat sich verändert. BERT bewertet die Ähnlichkeit mit 0,79. Für einen Menschen und in der Realität sind diese zwei Bausteine nicht identisch. Dagegen haben die Inhalte im medical outcome „reduction in MMD“ und „reduction in monthly migraine day (MMD)“ einen Ähnlichkeitsscore von 0,75. Dies ist immer noch ein hoher Wert jedoch niedriger als beim vorherigen Beispiel, obwohl diese aus menschlicher Sicht identisch sind (die Abkürzung wurde nur ausgeschrieben). Dies bekräftigt nochmal einen niedrigeren Threshold für medical outcome.

```
{
    "real_row": {
        "pubmed_id": 70707,
        "papername": "Advances in Rheumatoid Arthritis Therapy",
        "authors": "O'Connor, Sean; Walsh, Fiona",
        "publication_date": "2024-03-17",
        "journal": "Rheumatology International",
        "disease": "Arthritis",
        "medication_group": "No group specified",
        "therapy_or_medication": "Methotrexate",
        "medical_outcome": "Reduction in Pain",
        "result": 0.4,
        "outcome_category": "Symptom Intensity and Relief",
        "outcome_evaluation": "significantly improved"
    },
    "pred_row": {
        "id": 17,
        "pubmed_id": 70707,
        "paper": "Advances in Rheumatoid Arthritis Therapy",
        "authors": "O'Connor, Sean; Walsh, Fiona",
        "journal": "Rheumatology International",
        "publication_date": "2024-03-17",
        "disease": "Rheumatoid Arthritis",
        "medication_group": "Arthritis patients",
        "therapy_or_medication": "Methotrexate",
        "medical_outcome": "percentage reduction in pain levels",
        "result": "40%",
        "outcome_category": "Symptom Intensity and Relief",
        "outcome_evaluation": "significantly improved"
    },
    "best_similarity": {
        "disease_correct": 0.2728271484375,
        "medication_group_correct": 0.596625566482544,
        "therapy_or_medication_correct": 1.0,
        "medical_outcome_correct": 0.8682041168212891,
        "result_correct": 0.7082979679107666,
        "outcome_category_correct": 1.0,
        "outcome_evaluation_correct": 0.9999998211860657
    }
}
```

Abbildung 6-5: Ähnlichkeitswerte der Attribute zwischen einem realen und vorhergesagten Record

Tabelle 6-1: Accuracy für die einzelnen Attribute bei verschiedenen Thresholds (Alle Angaben in Prozent)

Threshold für Similarity	Complete	Disease	Medication Group	Therapy or Medication	Medical Outcome	Result	Outcome Category	Outcome Evaluation
0,1	100	100	100	100	100	100	100	100
0,2	99	100	100	100	100	99	100	100
0,25	99	100	100	100	100	99	100	100
0,3	95	96	100	100	100	99	100	100
0,4	86	94	93	100	100	99	100	100
0,5	51	94	56	99	100	99	100	100
0,6	38	94	41	99	99	98	100	97
0,7	10	94	19	97	87	72	85	92
0,75	4	92	19	91	85	37	83	92
0,8	1	91	18	86	71	37	71	87
0,85	1	82	16	84	64	34	71	87
0,9	1	76	12	82	57	32	71	85
0,95	1	75	12	79	44	31	71	85
1	0	62	6	49	22	20	64	7

Die Tabelle 6-1 zeigt Accuracy für verschiedene Attribute bei unterschiedlichen Thresholds der Ähnlichkeit (Similarity). Bei einem sehr niedrigen Threshold von 0,1 erreichen alle Attribute eine perfekte Genauigkeit von 100%. Mit steigendem Threshold nimmt die Genauigkeit jedoch deutlich ab. Bei einem Threshold von 0,5 fällt die Genauigkeit für "Complete" auf 51% und für "Medication Group" auf 56%, während "Disease" und "Medical Outcome" weiterhin eine hohe Genauigkeit von 94% bzw. 100% beibehalten. Zusammen mit „Disease“ bleiben die letzten zwei auch bei höheren Thresholds hoch. „Outcome Evaluation“ und „Outcome Category“ haben auch bei einem sehr hohen Threshold eine hohe Accuracy. Bei einem hohen Threshold von 1,0 sind die Genauigkeitswerte für die meisten Attribute stark reduziert, z.B. erreicht "Complete" nur noch 0% und "Therapy or Medication" 49%. „Disease“ und „Outcome Evaluation“ bleiben mit 62 % und 64 % dennoch hoch: Diese Ergebnisse sind eindeutig identisch und der Best Case bei dieser Evaluation.

Das Verhalten der Attribute spiegelt die vorher erwähnten Punkte mit dem Freiheitsgrad gut wider. Während Attribute mit Einschränkungen und kurzen, eindeutigen Antworten eine hohe Accuracy bei hohen Thresholds haben, fallen die Accuracy von Attribute mit hohen Freiheitsgraden schon bei kleineren Thresholds stark ab. Beispiele dafür sind Medication Group bei einem Threshold von 0,5 mit 93 % auf 56 % und 0,7 mit 41 % auf 19 % und Result bei 0,75 mit 72 % auf 37 %. Die Accuracy-Werte haben sich hier knapp halbiert bei einer Erhöhung des Thresholds um 0,1. „Unflexiblere“ Attribute haben einen Fall der Accuracy erst bei höheren Threshold. Diese Abstürze sind auch viel kleiner, die Ergebnisse sind stabiler.

Precision und Recall sind weitere entscheidende Metriken im Kontext der Accuracy, um die Qualität der Ergebnisse zu bewerten. Ein niedriger Threshold kann zu einem hohen Recall führen, da mehr relevante Ergebnisse erfasst werden, aber möglicherweise

auf Kosten der Precision, da auch irrelevante Ergebnisse eingeschlossen werden. Ein hoher Threshold kann die Precision verbessern, da hier wirklich nur die eindeutigsten Ergebnisse berücksichtigt werden, jedoch auf Kosten der Recall, da viele relevante Ergebnisse möglicherweise übersehen werden. Wie vorher beschrieben passiert dies auch. In der Tabelle zeigt sich dies darin, dass bei hohen Thresholds die Genauigkeit für einige Attribute stark abnimmt, was auf eine Abnahme der Recall hinweist, während die Precision für die verbleibenden Treffer möglicherweise höher ist.

Weiter muss erwähnt werden, dass eine Accuracy von 100 % unrealistisch ist, da nur 98 Records vorhergesagt wurden. 100 ergibt sich daraus, dass die realen Records mit allen Records verglichen wurden. So haben sie nicht passende Records als beste Vorhersage bekommen, bedeutet einige vorhergesagte Records kommen doppelt vor, da kein Äquivalent für das reale Record existiert.

Die Visualisierung der Ergebnisse mit den bestmöglichen Thresholds ist in Abbildung 6-6 zu sehen. Die Accuracy wird bestimmt durch die Anzahl der korrekten Vorhersagen geteilt durch die Anzahl der Records. Dies wird für jedes Attribut und Completely berechnet und als Balkendiagramm visualisiert. Da die Anzahl der Records 100 beträgt, ist die Accuracy einfach zu berechnen.

Folgende Thresholds wurden dabei ausgewählt:

- disease_correct: 0.8
- medication_group_correct: 0.6
- therapy_or_medication_correct: 0.75
- medical_outcome_correct: 0.75
- result_correct: 0.65
- outcome_category_correct: 0.9
- outcome_evaluation_correct: 0.9

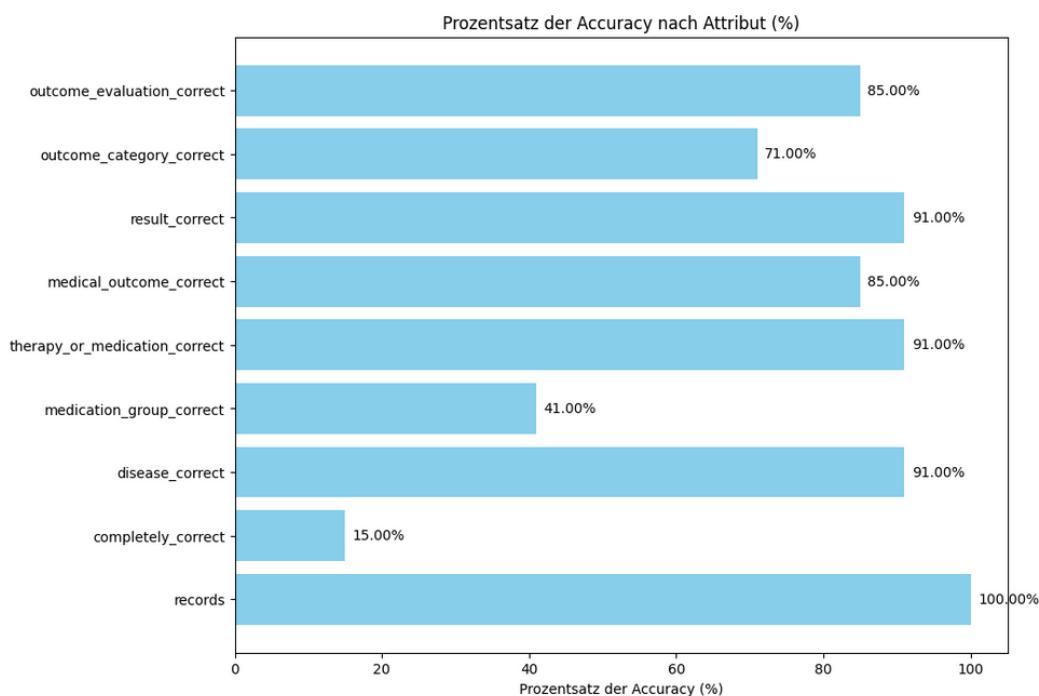


Abbildung 6-6: Balkendiagramm zur Visualisierung der Accuracy bei den passenden Thresholds

Das Diagramm zeigt den Prozentsatz der Accuracy für die verschiedenen Attribute bei den besten Thresholds. Die Genauigkeit für das Attribut „records“ ist nur ein Platzhalter und beinhaltet die Anzahl der realen Records.

Die Attribute „Disease“, „Medikation“ und „Result“ weisen hohe Accuracy von 91 % auf, was auf eine gute Identifikation dieser Werte in den vorhergesagten Medical Records hinweist. „Medical Outcome“ hat mit 85 % auch eine hohe Genauigkeit. Bei diesen Attributen kann ein hoher Recall und eine hohe Precision erwartet werden.

„Outcome Evaluation“ und „Outcome Category“ sind hier etwas besonders. Da beide einen hohen Threshold haben, sind die Inhalte mit hoher Sicherheit richtig identifiziert. Outcome Evaluation hat mit 85 % eine bessere Accuracy als Outcome Category mit 71 %. Grund dafür ist, dass es für Category mehr Kategorien bzw. Aufzählungen gibt (acht) als bei Evaluation (nur drei). So sind falsche Vorhersagen wahrscheinlicher. Bei diesen Attributen kann auch ein hoher Recall und eine hohe Precision erwartet werden.

„Medication Group“ hat von den Attributen die niedrigste Genauigkeit mit 41 %. Da der Threshold mit 0,6 am niedrigsten ist, ist davon auszugehen, dass die Precision ebenfalls nicht hoch ist. Das war zu erwarten, da Medication Group die meisten Freiheiten hat.

„Complete“ ist abhängig von der Accuracy der anderen Attribute und wird nur erhöht, wenn ein Record identisch ist, also wenn alle Attribute identisch sind. Damit kann Complete maximal die Accuracy von dem Attribut mit der niedrigsten Accuracy haben. Niedriger ist auch möglich, da sich die gleichen Inhalte der Attribute über die verschiedenen Records verteilen und nicht zwangsläufig in einem Record sind. Deswegen hat dieses Attribut die größten Sprünge und niedrigste Genauigkeit mit 15 %. Bedeutet nur 15 Records sind voraussichtlich komplett richtig, während bei voraussichtlich 85 % mindestens ein Fehler vorhanden ist. Durch die niedrige Genauigkeit von Medication Group war auch schon vorherzusehen, dass die Accuracy von Complete niedrig ist.

6.5 Abschließende Bewertung

Abschließend lässt sich sagen, dass die Performance des Modells auf dem ersten Blick gut aussieht. Die hohe Accuracy lassen auf eine Korrektheit der meisten Informationen schließen. Jedoch ist das nur ein Teil der Wahrheit. Einige unterschiedliche Inhalte werden falsch als identisch bewertet, da sie eine hohe Ähnlichkeit haben, während identische Inhalte als nicht ähnlich und somit falsch klassifiziert werden. Die Accuracy hier ist nur ein naiver Wert, die Dunkelziffer sieht anders aus und ist vermutlich sogar höher.

Manuelle Überprüfungen haben gezeigt, dass identische Werte eine niedrige Ähnlichkeit haben, weshalb sie als nicht identisch klassifiziert werden. Dies sind keine Einzelfälle. Über hohe Thresholds wurden viele Inhalte eliminiert, die zwar ähnlich jedoch nicht identisch sind. Hier wurde die Precision auf Kosten von Recall und Accuracy erhöht, was nochmal dafürspricht, dass die Genauigkeit höher sein kann, als hier berechnet.

Auch haben manuelle Überprüfungen gezeigt, dass die Ergebnisse selbst für eine unerfahrene Person weitestgehend verständlich sind. Hier wird aber noch ein Experte zur finalen Bewertung benötigt. Die Konsistenz der Ergebnisse schwankt. Wiederholte Durchläufe und Fragen geben nicht immer die gleichen Ergebnisse zurück. In Zukunft muss das Evaluations-Konzept verbessert und erweitert werden, um die Genauigkeit der Ergebnisse besser zu bewerten.

7 Ergebnisse und Diskussion

In diesem Kapitel werden die Ergebnisse, generierten Medical Records durch die Analyse und Abfragen der DB vorgestellt, analysiert und diskutiert. Dabei werden die Effektivität und Nutzungsmöglichkeiten der DB für eine Übersicht und (Meta-)Analysen betrachtet. Im Anhang D befinden sich weitere Ergebnisse und Abbildungen der Records und Sichten für verschiedene Anwendungsfälle und Abfragen, sowie alle SQL-Queries.

7.1 Einhaltung der Enumeration

Zuerst werden die Enumeration bzw. Aufzählungen Outcome Category und Outcome Evaluation und ihre Umsetzung und Einhaltung LLaMa3 betrachtet.

7.1.1 Outcome Category

Zwar wurde LLaMa3 im Prompt gesagt, sich nur auf die folgenden Aufzählungen einzuschränken, jedoch hat es für manche Medical Records neue Kategorien erfunden oder Beispiele mit aufgenommen. Nicht immer wurde diese Eingrenzung eingalten.

- Efficacy Measures
- Symptom Intensity and Relief
- Adverse Events and Side Effects
- Quality of Life
- Physiological Measure
- Medication Usage
- Health Economics
- No category

Werden die Medical Records nach den Kategorien gruppiert, machen diese Aufzählungen trotzdem 97,95 % aller Datensätze aus, also insgesamt ca. 3.584 Records von 3.659. Die erfundenen Kategorien kommen nur vereinzelt vor (siehe Abbildung 7-1).

Die gleichen Kategorien, die jedoch Beispiele dabeihaben, die so auch im Prompt zu finden sind, machen 36 Records aus und damit 0,98 %. Werden sie korrigiert, sind insgesamt 3.620 und 98,93 % richtig klassifiziert. Ein sehr gutes Ergebnis. LLaMa3 hat diese Enumeration sehr gut eingehalten und umgesetzt.

7.1.2 Enumeration – Outcome Evaluation

Auch hier wurde LLaMa3 gesagt, sich nur auf die folgenden Aufzählungen einzuschränken, jedoch hat es neue Kategorien erfunden oder Beispiele mit aufgenommen.

- Significantly worse
- Significantly improved
- No significant difference

Zusammen machen diese Aufzählungen trotzdem 92,71 % aller Datensätze aus, also insgesamt 3.392 Records von 3.659. Die erfundenen Kategorien kommen nur vereinzelt vor und ähneln sich sehr den originalen Kategorien (siehe Abbildung 7-2).

Interessanterweise haben sich auch 50 Rechtschreibfehler ergeben: 44 mit „significantly improved“ und sechs mit „significantly worse“. Auf das Prompt ist es nicht zurückzuführen, da es dort richtig steht. Weiter gibt es auch sechs Einträge, die identisch

mit „no significant difference“ sind, jedoch case sensitive sind. Sie haben No statt no und werden deshalb als eigene Kategorie gezählt.

Werden diese 56 Einträge dazugezählt beträgt die Anzahl 3.448 Records und der Anteil 94,23 %. Das sind 1,5 % mehr, wenn diese kleinen Fehler ausgebügelt werden.

	outcome_category character varying (255)	counter bigint	share_percentage numeric
1	Efficacy Measures	1868	51.05
2	Physiological Measure	576	15.74
3	Adverse Events and Side Effects	395	10.80
4	Symptom Intensity and Relief	372	10.17
5	Medication Usage	148	4.04
6	Quality of Life	120	3.28
7	No Category	53	1.45
8	Health Economics	52	1.42
9	Efficacy Measures (such as responder rates)	17	0.46
10	Physiological Measure (such as heart rate)	11	0.30
11	Medication Usage (such as number of days in medication)	8	0.22
12	Health Care Utilization	4	0.11
13	Patient satisfaction	3	0.08
14	Pharmacokinetics	3	0.08
15	Portal inflammation improvement	1	0.03
16	Treatment Patterns	1	0.03
17	Risk of CM onset	1	0.03
18	Patient-reported improvement	1	0.03

Total rows: 43 of 43 Query complete 00:00:00.080

Abbildung 7-1: Anzahl der Records nach Outcome Category für eine Teilmenge

	outcome_evaluation character varying (255)	counter bigint	share_percentage numeric
1	significantly improved	2107	57.58
2	significantly worse	658	17.98
3	no significant difference	627	17.14
4	significantly better	105	2.87
5	signifcantly improved	44	1.20
6	significantly lower	24	0.66
7	significantly reduced	16	0.44
8	significantly higher	10	0.27
9	significantly decreased	6	0.16
10	No significant difference	6	0.16
11	signifcantly worse	6	0.16
12	significantly increased	4	0.11
13	significantly greater	3	0.08
14	significantly changed	2	0.05

Total rows: 48 of 48 Query complete 00:00:00.079

Abbildung 7-2: Anzahl der Records nach Outcome Evaluation

7.2 Die Records

Die nachfolgende Abbildung zeigt die View auf die Daten als Tabelle:

	medication_group character varying (255)	therapy_or_medication character varying (255)	medical_outcome character varying (255)	result character varying (255)	outcome_category character varying (255)	outcome_evaluation character varying (255)
1	Randomised group	Erenumab	discontinuation due to adverse events	10.6%	Adverse Events and Side Effects	significantly worse
2	Randomised group	Erenumab	≥50% reduction from baseline in monthly migraine days	55.4%	Efficacy Measures	significantly improved
3	Randomised group	Topiramate	discontinuation due to adverse events	38.9%	Adverse Events and Side Effects	significantly worse
4	Randomised group	Topiramate	≥50% reduction from baseline in monthly migraine days	31.2%	Efficacy Measures	significantly better
5	No group specified	OnabotulinumtoxinA (BoN...)	Responder rate	59.1%	Efficacy Measures	significantly improved
6	No group specified	OnabotulinumtoxinA (BoN...)	Mean reduction in MMDs	9.4 days	Efficacy Measures	significantly improved
7	No group specified	OnabotulinumtoxinA (BoN...)	Mean reduction in MHDs	14.9 days	Efficacy Measures	significantly improved
8	Participants with migraine	Erenumab	≥50% reduction in monthly migraine days	61%	Efficacy Measures	significantly improved
9	CGRP-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	60 of 99 (61%)	Efficacy Measures	significantly improved
10	CGRP-non-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	13 of 25 (52%)	Efficacy Measures	no significant difference
11	Participants with EM and 2-4 prior ...	Erenumab (140 mg)	≥50% reduction in monthly migraine days (MMDs)	79/151 participants (52.3%)	Efficacy Measures	significantly improved
12	Continuous erenumab group	Erenumab (140 mg)	≥50% responders in at least half of OLEP visits	26/35 (74.3%)	Efficacy Measures	significantly improved
13	Placebo-erenumab group	Erenumab (140 mg)	≥50% responders in at least half of OLEP visits aft...	42/103 (40.8%)	Efficacy Measures	significantly improved
14	Overall population	Erenumab (140 mg)	Mean MMD change from baseline	-4.4 [3.9] days	Efficacy Measures	significantly improved
15	Japanese patients	Erenumab	Changes from baseline in monthly migraine days ...	-3.8 (0.4) days	Efficacy Measures	significantly improved
16	Japanese patients	Erenumab	Changes from baseline in monthly acute migraine days	-2.6 (0.4) days	Medication Usage (such as num...	significantly improved
17	Japanese patients	Erenumab	Proportion of ≥50% responders for MMD reduction	34.1%	Efficacy Measures	significantly improved
18	Japanese patients	Erenumab	Proportion of ≥50% responders for MMD reduction	44.4%	Efficacy Measures	significantly improved
19	Japanese patients	Erenumab	Exposure-adjusted incidence of treatment-emergent ...	219.7 per 100 patient-years	Adverse Events and Side Effects	significantly worse
20	Patients with chronic migraine (C...	Erenumab	Reversion from CM to EM	52.3%	Efficacy Measures	significantly improved
21	No group specified	Erenumab	Discontinuation of headache days	55.4%	Efficacy Measures	significantly improved
22	Patients with high frequency episodic ...	Anti-CGRP mAbs (galcanezumab)	MIDAS scores	Significantly lower	Quality of Life	significantly improved
23	Patients with high frequency episodic ...	Anti-CGRP-R mAbs (erenumab)	MIDAS scores	Comparable	Quality of Life	significantly improved
24	Patients with high frequency episodic ...	Anti-CGRP mAbs (galcanezumab)	Number of mean migraine days	Significantly lower	Efficacy Measures	significantly improved
25	Patients with high frequency episodic ...	Anti-CGRP-R mAbs (erenumab)	Number of mean migraine days	Comparable	Quality of Life	significantly improved
26	140 patients at Headache Centre o...	Galcanezumab, Fremenezumab	Mean reduction of migraine monthly days from baseline	-12.0 (-9.8, -14.1)	Efficacy Measures	significantly improved

Abbildung 7-3: Schnelle Übersicht der strukturierten Medical Records Daten in der PubMedMA-DB

Aus dieser Sicht bzw. View können viele Informationen und viel Wissen generiert werden für verschiedene Analysen. Es ist zu sehen welche Medikamente welche Wirkung bei welcher Krankheit haben. Teilweise steht auch in den Records auf welche Patientengruppe es sich bezieht. Über die Kategorie wird bestimmt welcher Bereich durch das Medikament verändert wird und in welchem Ausmaß. Jedoch gibt es mit 3.659 Medical Records eine große Menge an Daten und medizinischen Studien. Für eine geordnete und gezielte Analyse muss nach bestimmten Daten, Gruppen und kleineren Datensätzen gesucht werden.

Hierbei kommt die Stärke der PostgreSQL-DB zum Einsatz. Durch die strukturierte Sammlung und den vielen SQL-Funktionen zum Filtern, Aggregieren und Gruppieren von Daten wird es vereinfacht die notwendigen Daten abzufragen, diese vorzubereiten und zu analysieren. In den nächsten Abschnitten werden die Medical Records in der DB für verschiedene Abfrage näher betrachtet, die für Übersichten und Meta-Analysen verwendet werden können.

7.2.1 Doppelte Records

Es kann vereinzelt vorkommen, dass ein vom Inhalt her identischer Record zweimal vorkommt, wie in Abbildung 7-4 zu sehen ist. Beide Records unterscheiden sich diesbezüglich in den Resultaten. Bei beiden wird ein Resultat von 61 % angegeben, wobei im ersten Record auch die Anzahl der Patienten angegeben wird mit 60 von 99. Da beide sich auf das gleiche Paper bzw. die gleiche Studie beziehen, kann es vorkommen, dass die Analyse verfälscht wird, da dieser gute Wert zwei Mal miteinfließt, obwohl dieses Ergebnis nur einmal vorkam. Der Record mit der Angabe der Patientenzahl sollte behalten werden, da dieses Record mehr Informationen aufzeigt. Im Bestfall sollte die Anzahl im Attribut Medication Group zu finden sein.

≥50% reduction in monthly migraine days	60 of 99 (61%)	significantly improved
≥50% reduction in monthly migraine days	61%	significantly improved

Abbildung 7-4: Duplikat eines Medical Records mit unterschiedlich repräsentierten Resultaten

7.2.2 Anzahl der Records pro Gruppe

Um sich eine Übersicht zu verschaffen, wie viele Paper medizinische Studien über die gleiche Medikation gemacht haben, können die Medical Records nach Medikation gruppiert und die Medical Records gezählt werden. So kann wie in Abbildung 7-5 herausgefunden werden, dass es für Erenumab 79 Medical Records aus verschiedenen Paper gibt. Insgesamt gibt es in der PubMedMA-DB 1500 Medikamente und Therapien, wobei die meisten nur ein Record haben.

	therapy_or_medication character varying (255)	count bigint
1	Erenumab	79
2	Eptinezumab	76
3	Fremanezumab	70
4	Lasmiditan	64
5	Galcanezumab	56
6	Topiramate	54
7	Duloxetine	49

Total rows: 1500 of 1500 | Query complete 00:00:00.156

Abbildung 7-5: Anzahl der Medical Records nach Medikament

Die Gruppierung der Daten kann um einen Schritt erweitert werden. Zusätzlich kann die Krankheit miteinbezogen werden, um so zu erfahren, wie viele Paper es für ein Medikament für eine spezielle Krankheit gibt. So können die Ergebnisse mehrerer Studien gefiltert und zusammengefasst werden, um daraus aussagekräftige Ergebnisse für eine Fragestellung zu berechnen. Abbildung 7-6 zeigt diese Aggregation. Hier ist zu sehen, dass Erenumab, Lasmiditan und Fremanezumab im Kontext von Migräne alle 59 Paper haben. Insgesamt gibt 1.909 Zeilen, wobei die meisten wieder nur ein Record haben.

	disease character varying (255)	therapy_or_medication character varying (255)	count bigint
1	Migraine	Fremanezumab	59
2	Migraine	Lasmiditan	59
3	Migraine	Erenumab	59
4	Migraine	Eptinezumab	52
5	Migraine	Galcanezumab	41
6	Migraine	Rimegepant	31
7	Migraine	Ubrogepant	30
8	Migraine	Eletriptan	24
9	Migraine	Atogepant	21
10	Migraine	Placebo	17

Total rows: 1909 of 1909 | Query complete 00:00:00.156

Abbildung 7-6: Anzahl der Medical Records nach Medikament

Auch kann nach der Krankheit kann die die Anzahl der Paper angefragt werden. So können medizinische Institute erfahren welche Krankheiten schon viele Paper und

Studien haben und welche nicht. Bei den Krankheiten mit wenig Paper gibt es vielleicht eine Forschungslücke, die so über die DB entdeckt wird. Es kann zu Anstrengungen in diesem Bereich kommen. Abbildung 7-7 zeigt so eine View. Das Migräne hier ganz oben ist, war vorherzusehen, da dieses Projekt diese Krankheit als Hauptfokus gewählt hat. Interessanterweise kommen die anderen gesuchten Krankheit wie Bronchitis (500 Paper wurden angefragt) nicht in den höchsten 10 vor.

	disease character varying (255)	count bigint
1	Migraine	1166
2	Chronic Migraine	161
3	Epilepsy	49
4	Depression	36
5	Major Depressive Disorder (MDD)	31
6	Cluster headache	25
7	Status Migrainosus	24
8	Episodic Migraine	23

Abbildung 7-7: Anzahl der Paper nach Krankheit

7.2.3 Records für ein Medikament

In Abbildung 7-8 werden die Medical Records für Erenumab und Migräne gefiltert und ausgegeben. Hier sind verschiedene Tests und Resultate aufgelistet aus verschiedenen Paper. In den ersten vier Zeilen aus zwei Paper ist zu sehen, dass Erenumab bei Migräne dafür sorgt, dass die MMD um mindestens 50 % sinkt bei zwischen 34,1 % und 61 % der Patienten. Diese Werte sind hier sinnvoll und können dafür genutzt werden, um zu analysieren wie viele Patienten im Durchschnitt eine Reduktion der MMD um mindestens 50 % haben.

Abbildung 7-9 zeigt noch mehr Medical Records zu Erenumab und zu MMD. Diese Sicht kann für eine Meta-Analyse verwendet werden, in der die Fragestellung „Welche Einflüsse hat Erenumab auf MMD?“ analysiert wird. Hier werden viele Paper abgedeckt mit verschiedenen Patientengruppen, Medical Outcomes bzw. Ergebnissen und Resultaten. Die Evaluation des Outcomes zeigt zusätzlich auch, dass sich in den meisten Fällen die Efficacy Measures signifikant verbessert hat. Es gab also eine signifikante Reduktion der Krankheit.

pubmed_id	disease	medication_group	therapy_or_medicament	medical_outcome	result	outcome_category	outcome_evaluation
1	34743579	Migraine	Randomised group	Erenumab	≥50% reduction from baseline in monthly migraine days	55.4%	Efficacy Measures
2	38859744	Migraine	Participants with migraine	Erenumab	≥50% reduction in monthly migraine days	61%	Efficacy Measures
3	38859744	Migraine	CGRP-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	60 of 99 (61%)	Efficacy Measures
4	38859744	Migraine	CGRP-non-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	13 of 25 (52%)	Efficacy Measures
5	38783593	Migraine	Patients with chronic migraine (CM) from...	Erenumab	Reversion from CM to EM	52.3%	Efficacy Measures
6	38783593	Migraine	No group specified	Erenumab	Discontinuation of headache days	55.4%	Efficacy Measures
7	37940860	Migraine	No group specified	Erenumab	reversion from chronic migraine to episodic migraine	57.3%	Efficacy Measures
8	37940860	Migraine	No group specified	Erenumab	transition from medication overuse to non-overuse	46.5%	Efficacy Measures
9	38823758	Migraine	No group specified	Erenumab	MMD	-7.18 days	Efficacy Measures
10	38823758	Migraine	No group specified	Erenumab	HIT-6	-6.97	Efficacy Measures
11	38823758	Migraine	No group specified	Erenumab	NDM	-6.22 days	Efficacy Measures
12	37940860	Migraine	Adult patients included in the FHU InovP...	Erenumab	effectiveness and safety assessment	74/140 (52.9%) responders at M3	Efficacy Measures
13	37940860	Migraine	No group specified	Erenumab	monthly migraine days	18.0 (13.0-26.0)	Physiological Measure
14	38823758	Migraine	No group specified	Erenumab	AMI	-15.75	Efficacy Measures
15	38823758	Migraine	No group specified	Erenumab	PI	-1.71	Efficacy Measures
16	38823758	Migraine	Approximately one-third of patients	Erenumab	MMD/MHD response rate	Response greater than 30%	Efficacy Measures
17	38823758	Migraine	Approximately one-sixth of patients	Erenumab	MMD/MHD response rate	Response exceeding 50%	Efficacy Measures
18	38823758	Migraine	3-4% of patients	Erenumab	MMD/MHD response rate	Response rate of 100%	Efficacy Measures
19	38823758	Migraine	No group specified	Erenumab	Adverse events	0.34 at 6-months	Adverse Events and Side Effects
20	38823758	Migraine	No group specified	Erenumab	Adverse events	0.43 at 12-months	Adverse Events and Side Effects

Abbildung 7-8: Die Medical Records zu Migraine und Erenumab

pubmed_id	disease	medication_group	therapy_or_med	medical_outcome	result	outcome_category	outcome_evaluation
integer	character varying (25)	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)
34743579	Migraine	Randomised group	Erenumab	≥50% reduction from baseline in monthly migraine days	55.4%	Efficacy Measures	significantly improved
38859744	Migraine	Participants with migraine	Erenumab	≥50% reduction in monthly migraine days	61%	Efficacy Measures	significantly improved
38859744	Migraine	CGRP-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	60 of 99 (61%)	Efficacy Measures	significantly improved
38859744	Migraine	CGRP-non-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	13 of 25 (52%)	Efficacy Measures	no significant difference
37597868	Episodic Migraine...	Japanese patients	Erenumab	Changes from baseline in monthly migraine days (MMD)	-3.8 (0.4) days	Efficacy Measures	significantly improved
37597868	Episodic Migraine...	Japanese patients	Erenumab	Proportion of ≥50% responders for MMD reduction	34.1%	Efficacy Measures	significantly improved
37597868	Episodic Migraine...	Japanese patients	Erenumab	Proportion of ≥50% responders for MMD reduction	44.4%	Efficacy Measures	significantly improved
38823758	Migraine	No group specified	Erenumab	MMD	-7.18 days	Efficacy Measures	significantly improved
37940860	Migraine	No group specified	Erenumab	monthly migraine days	18.0 (13.0-26.0)	Physiological Measure	significantly improved
38823758	Migraine	Approximately one-third of patients	Erenumab	MMD/MHD response rate	Response greater than 30%	Efficacy Measures	significantly improved
38823758	Migraine	Approximately one-sixth of patients	Erenumab	MMD/MHD response rate	Response exceeding 50%	Efficacy Measures	significantly improved
38823758	Migraine	3-4% of patients	Erenumab	MMD/MHD response rate	Response rate of 100%	Efficacy Measures	significantly improved
37498837	Migraine	Japanese patients with different migr...	Erenumab	change from baseline in MMD	similar across LFM and HFEM/CM subgroups	Efficacy Measures	similar
37498837	Migraine	Japanese patients with different migr...	Erenumab	proportion of patients achieving at least 50% or 75% reduction from baseline in MMD and MS...	similar across migraine frequency groups	Efficacy Measures	no significant difference
34021585	Migraine	No group specified	Erenumab	Mean difference in MMD change from baseline	-0.06 [-0.61, 0.50]	Physiological Measure (such as heart rate)	significantly improved
38941460	Chronic Migraine ...	Patients with CM with more than two tr...	Erenumab	Reductions in Monthly Migraine Days (MMD)	568 MDs	Efficacy Measures	significantly improved
38491460	Episodic Migraine...	Patients with EM with one or more prior...	Erenumab	Reductions in Monthly Migraine Days (MMD)	172 MDs	Efficacy Measures	significantly improved

Abbildung 7-9: Medical Records zu Erenumab

Abbildung 7-10 und Abbildung 7-11 zeigen die Medical Outcomes und Resultate zu MIDAS-Score und HIT-6 Score an für verschiedene Medikamente. Diese geben sinnvolle Werte zurück, die vergleichbar sind innerhalb des Medikaments. Als Beispiele gibt es hierfür die ersten zwei Zeilen und die achte Zeile in Abbildung 7-10 für Erenumab (-8.3 bis -4.8, -6.97 und -11.1 points) sowie die zehnte und elfte Zeile für Fremanezumab (-16.9 und 17.9 points). Mit dieser Übersicht können die Unterschiede zwischen den Medikamenten analysiert werden.

Es ist auch zu sehen das die Resultate in verschiedenen Formen präsentiert werden, so gibt es neben den hier genannten points auch absolute Zahlen mit weiteren Werten, Tage und Wahrscheinlichkeiten. Diesen Fall gibt es auch bei Krankheiten und Medikationen. Bei Krankheiten kann es vorkommen, dass die Abkürzung mitgeschrieben oder der erste Buchstabe klein geschrieben wird. Ähnlich ist es bei den Medikamenten. Hier kann zusätzlich noch die Dosis vorkommen. Das führt zu dem Problem, dass eine Gruppierung erschwert wird, da es eigenständige Gruppen sind (Migraine und migraine sind verschiedene Gruppen, sowie Chronic Migraine und Chronic Migraine (CM) als auch Erenumab und Erenumab 20mg). Das erschwert die Analysen. Hier ist eine Standardisierung der Einträge notwendig, die dafür sorgt, dass gleiche Daten auch gleich präsentiert werden.

Bei einigen Resultaten stehen statt numerischen Werten auch nur Texte wie „Significantly lower“. Die Suche nach den Paper hat gezeigt, dass sich in den Resultaten tatsächlich keine Werte befinden, sondern die Ergebnisse mit Text beschrieben werden. Dieses Problem ist somit nicht auf die Analyse durch LLaMa3 zurückzuführen, sondern auf das Abstract.

pubmed_id	disease	medication_group	therapy_or_medication	medical_outcome	result	outcome_category	outcome_evaluation
integer	character varying (25)	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)
33421995	Chronic Migraine	Patients with difficult-to-control chronic migraine	Erenumab	HIT-6 scores	-8.3 to -4.8, p<0.001	Medication Usage	significantly improved
38823758	Migraine	No group specified	Erenumab	HIT-6	-6.97	Efficacy Measures	significantly improved
38990457	Resistant migraine...	No group specified	Anti-CGRP mAbs (erenumab, galcanezumab, fremanezumab)	Reductions in MIDAS, HIT-6, MMD and MAM from baseline to ...	N/A	Quality of Life	significantly improved
38992666	Chronic Migraine...	No group specified	OnabotulinumtoxinA	Mean changes from baseline on HIT-6 questionnaire	-6.5	Symptom Intensity and Relief	significantly improved
38887386	Episodic Migraine...	Overall study population	Galcanezumab injections	HIT-6 scores	<-0.001	Quality of Life	significantly improved
38753586	Migraine	Asian patients with chronic migraine	Eptinezumab	Change from baseline in Headache Impact Test-6 (HIT-6) score	-24.6 points (95% CI 2.82-46.38; p = 0.028)	Quality of Life	significantly improved
38757125	Migraine	Patients with migraine	Selenium supplementation	Headache frequency, severity, HIT-6 score	-8.15 ± 0.77 vs -4.12 ± 0.77 (p < 0.001)	Symptom Intensity and Relief	significantly improved
38668603	Chronic Migraine	Patients with chronic migraine	anti-CGRP mAbs	mean reduction in HIT-6	-11.1 points	Efficacy Measures	significantly improved
38668603	Chronic Migraine	Patients with chronic migraine	BOT-NTA	mean reduction in HIT-6	-3.2 points	Efficacy Measure	significantly worse
38451463	Migraine	High-frequency episodic migraine (HfEM) or chronic ...	Fremanezumab	Reduction in HIT-6 score	-16.9	Efficacy Measures	significantly improved
38451463	Migraine	Chronic migraine (CM) with > 3 treatment failures	Fremanezumab	Reduction in HIT-6 score	-17.9	Efficacy Measures	significantly improved
38398444	Migraine	No group specified	Erenumab, Galcanezumab, Fremanezumab	Change from baseline in MIDAS, H-6, MMDs and MAD scores	Statistically significant	Quality of Life	significantly improved
38238659	Migraine	68 patients (85.3% female, mean age: 46.2 ± 13.1 year...)	CGRP mAb (galcanezumab, fremanezumab, and erenumab)	reduction in monthly migraine days (MMDs) and HIT-6 score	7.4 ± 5.6 (p < 0.0001), 56.2 ± 8.7 (P < 0.0001)	Efficacy Measures	significantly improved

Abbildung 7-10: Betrachtung des HIT-6 Score und des Outcomes für verschiedene Medikamente

pubmed_id	disease	medication_group	therapy_or_medication	medical_outcome	result	outcome_category	outcome_evaluation
integer	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)	character varying (255)
36856538	Migraine	Patients with high frequency episodic and chronic migraine	Anti-CGRP mAb (galcanezumab, fremanezumab)	MIDAS scores	Significantly lower	Quality of Life	significantly improved
36856538	Migraine	Patients with high frequency episodic and chronic migraine	Anti-CGRP-R mAb (erezumab)	MIDAS scores	Comparable	Quality of Life	significantly improved
38988460	Chronic Migraine	Patients with chronic migraine	Erenumab and Onabotulinumtoxin A (onabotNTA)	MIDAS score assessed 3 months after introduction of erenum...	Significant improvement	Quality of Life	significantly improved
38988460	Chronic Migraine	No group specified	Erenumab and Onabotulinumtoxin A (onabotNTA) monotherapy vs. initiation of onabotNTA-a monother...	MIDAS score reduction comparison	Significantly greater reduction	Quality of Life	significantly improved
38518587	Migraine	No group specified	Anti-CGRP mAb	MIDAS score improvement	Significant improvement	Quality of Life	significantly improved
38308209	Chronic migraine	No group specified	Anti-CGRP mAb	MIDAS reduction	-31.7 points	Quality of Life	significantly improved
38308209	Chronic migraine	No group specified	BOT-NTA	MIDAS reduction	-19.2 points	Quality of Life	significantly improved
35804294	Migraine	Patients with chronic migraine (CM)	Eptinezumab	MIDAS quantified disability	56.7 days reduction	Quality of Life	significantly improved
35804294	Migraine	Patients with very severe MIDAS disability	Eptinezumab	MIDAS quantified disability	32 days reduction at WK104	Quality of Life	significantly improved
37221478	Migraine	All patients	Fremanezumab	MIDAS scores decreased	23.0 (73.3 ± 56.8 to 50.3 ± 52.9)	Quality of Life	significantly improved

Abbildung 7-11: Betrachtung des MIDAS Score und des Outcomes für verschiedene Medikamente

7.2.4 Records zum Vergleich von zwei verschiedenen Medikamenten

Eine weitere Möglichkeit ist der Vergleich zwischen zwei Medikamenten für eine Krankheit für ein bestimmtes Outcome (wie im Beispiel von Abbildung 7-12 und Abbildung 7-13). Bei der ersten View werden die Outcomes auf die MMD (meist die Reduktion) der beiden betrachtet. Erenumab bietet hier im Schnitt eine etwas bessere Reduktion prozentual gesehen als Topiramate. Dies kann dazu führen, dass mehr Forschung in Richtung Erenumab gemacht wird und Unternehmen und Ärzte eher auf dieses Medikament setzen. Eventuell kann es auch mehr Forschung Richtung Topiramate geben, um die Schwächen und Probleme zu erkennen und die Wirkungen zu verbessern.

disease	medication_group	therapy_or_medication	medical_outcome	result	outcome_category	outcome_evaluation
Migraine	Randomised group	Erenumab	≥50% reduction from baseline in monthly migraine days	55.4%	Efficacy Measures	significantly improved
Migraine	Participants with migraine	Erenumab	≥50% reduction in monthly migraine days	61%	Efficacy Measures	significantly improved
Migraine	CGRP-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	60 of 99 (61%)	Efficacy Measures	significantly improved
Migraine	CGRP-non-sensitive participants	Erenumab	≥50% reduction in monthly migraine days	13 of 25 (52%)	Efficacy Measures	no significant difference
Episodic Migraine (E...	Japanese patients	Erenumab	Changes from baseline in monthly migraine days (MMD)	-3.8 (0.4) days	Efficacy Measures	significantly improved
Migraine	No group specified	Erenumab	monthly migraine days	18.0 (13.0-26.0)	Physiological Measure	significantly improved
Migraine	Randomised group	Topiramate	≥50% reduction from baseline in monthly migraine days	31.2%	Efficacy Measures	significantly better
Episodic Migraine	Adults diagnosed with episodic migraine	Topiramate	Reduction of monthly migraine days (MMDs)	-1.11	Efficacy Measures	no significant difference
Chronic migraine	No group specified	Topiramate	50% reduction in monthly migraine days	50.06	Efficacy Measures	significantly improved
Chronic Migraine	No group specified	Topiramate	mean change in monthly migraine days	-3.25	Efficacy Measures	significantly improved
Migraine	No group specified	Topiramate	monthly migraine days (MMD) or moderate to severe headache da...	-1.45 (-1.52 to -1.38)	Efficacy Measures	significantly worse

Abbildung 7-12: Vergleich zwischen zwei Medikamenten für eine Krankheit und ein Medical Outcome

Die zweite View zeigt einen Vergleich der Outcome Kategorien und Evaluationen an. Diese Records können dann dementsprechend abgefragt und analysiert werden. Hier ist zu sehen, das es deutlich mehr Paper zu Erenumab gibt und in deutlich mehr Studien gezeigt wurden, dass die Efficacy Measures sich durch Erenumab signifikant verbessert.

Allgemein können die zwei Attribute mit Kategorien verwendet werden, um gezielt nach Veränderungen in einem bestimmten Bereich (wie Efficacy Measures oder Health Economics) zu suchen und die dazugehörigen Records zu analysieren (z. B. für ein bestimmtes Medikament). So können alle Records gefiltert werden, die eine signifikante Verschlechterung bei Physiological Measure haben, um dadurch herauszufinden, was der Grund dafür ist und welche Medikamente oder Therapien betroffen sind.

therapy_or_medication	outcome_category	outcome_evaluation	count
Topiramate	Adverse Events and Side Effects	significantly improved	1
Topiramate	Adverse Events and Side Effects	significantly worse	2
Topiramate	Efficacy Measures	significantly better	2
Topiramate	Efficacy Measures	significantly improved	2
Topiramate	Efficacy Measures	significantly worse	2
Erenumab	Adverse Events and Side Effects	no significant difference	2
Erenumab	Adverse Events and Side Effects	significantly improved	1
Erenumab	Adverse Events and Side Effects	significantly worse	3
Erenumab	Efficacy Measures	no significant difference	2
Erenumab	Efficacy Measures	significantly improved	4
Erenumab	Efficacy Measures	significantly improved	31
Erenumab	Efficacy Measures	similar	1
Erenumab	Medication Usage	significantly worse	2
Erenumab	No Category	no significant difference	1

Abbildung 7-13: Vergleich von zwei Medikamenten in den Outcome Kategorien und Evaluationen

7.2.5 Records nach Jahr und Medikament

Die Abfragen können auch genutzt werden, um zu erfahren wie viele Paper für ein Medikament oder für eine Krankheit veröffentlicht wurden. So können unter anderem Trends erkannt werden, welche Medikamente gerade mehr und mehr genutzt werden und welche weniger. Abbildung 7-14 und Abbildung 7-15 zeigt dies für zwei Medikamente. Bei der ersten Abbildung zu Erenumab ist zu sehen, dass die Anzahl der Paper sich über die Jahre seit 2022 stark erhöht hat (von 4 auf 12 in 2024), wobei das Jahr 2024 noch nicht vorbei ist, was theoretisch auf einen Trend hinweisen könnte. Ein anderer Fall ist bei Lasmiditan zu sehen. Hier nahm bis 2022 die Anzahl der Paper um 50 % auf 12 zu im Vergleich zum Vorjahr, jedoch nahm diese Menge im nächsten Jahr ab. Es wurden wieder nur 8 Paper veröffentlicht. Im Jahr 2024 ist eine geringere Menge zu vermuten, was einen negativen Trend aufzeigt.

Natürlich entspricht das hier nicht unbedingt der Wahrheit, jedoch könnte bei einer sehr großen Datenmenge eine fundierte Trend-Analyse gemacht werden.

Erenumab	2024	12
Erenumab	2023	11
Erenumab	2022	3
Erenumab	2021	4

Abbildung 7-14: Anzahl der veröffentlichten Paper nach Jahr für Erenumab

Lasmiditan	2024	2
Lasmiditan	2023	8
Lasmiditan	2022	12
Lasmiditan	2021	8

Abbildung 7-15: Anzahl der veröffentlichten Paper nach Jahr für Lasmiditan

Eine Abfrage nach der Gesamtanzahl der veröffentlichten Paper pro Jahr zeigt, dass deutlich mehr Paper in 2024 veröffentlicht wurden. Eventuell gibt hier PubMed eher aktuelle Paper bei Anfragen zurück oder die Anzahl an veröffentlichten Paper hat stark zugenommen durch die Verfügbarkeit von bestimmten Werkzeugen. Leider hat es auch zu bedeuten, dass die Zustände in den Abfragen nicht ganz der Realität entsprechen. Für eine klare Trenderkennung werden dringend viel mehr Paper benötigt.

7.2.6 Schwankende und schwindende Records

Das nachfolgende Bild zeigt verschiedene Records aus einem Paper, die so nicht immer erkannt werden. Ein Lauf garantiert nicht, dass die gleichen Records generiert werden, wodurch es Schwankungen in der Anzahl an Records gibt pro Durchlauf und Record verschwinden.

Die Records in Tabelle 7-1 und Tabelle 7-2 zeigen Records aus einem alten Run, die es in den aktuellen Daten nicht mehr gibt und die in dem aktuellen Run nicht mehr generiert wurden. Diese zeigen nochmal sinnvolle Ergebnisse für den HIT-6 und MIDAS Score für Erenumab. Für die Zeile mit PubMed ID 33421995 in Tabelle 7-1 gab es als Feedback, dass hier eher die Anzahl der red days, green days usw. als Outcome und Result benötigt werden. Diese gab es ihm alten Durchlauf nicht, jedoch gibt es sie im jetzigen (siehe Abbildung 7-16). Das zeigt nochmal, wie die Ergebnisse und Genauigkeit schwanken können.

pubmed_id	disease	medication_group	therapy_or_med	medical_outcome	result	outcome_category	outcome_evaluation
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	Number of red days	-6.4 days (SE 0.67, 95%CI -7.7 to -5.1, p=0.001)	Efficacy Measures	significantly improved
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	Number of green days	-6.8 days (SE 0.96, 95%CI -8.80 to -4.9, p=0.001)	Efficacy Measures	significantly improved
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	Number of triptan days	-6.5 days (SE 0.86, 95%CI -8.3 to -4.8, p=0.001)	Medication Usage	significantly improved
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	Number of painkiller days	-6.5 days (SE 0.86, 95%CI -8.3 to -4.8, p=0.001)	Medication Usage	significantly improved
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	HIT-6 scores	-8.3 to -4.8, p=0.001	Medication Usage	significantly improved
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	PHQ-9 scores	-8.3 to -4.8, p=0.001	Medication Usage	significantly improved
33421995	Chronic Migraine	Patients with difficult-to-...	Erenumab	PDI scores	-8.3 to -4.8, p=0.001	Efficacy Measures	significantly improved

Abbildung 7-16: Erkennung von komplexen Records, die nicht bei jedem Lauf erkannt werden

Tabelle 7-1: Alte Records für HIT-6 Scores für das Medikament Erenumab

PubMed ID	Publikationsdatum	Krankheit	Medikament	Medical Outcome	Result
38823758	02.06.2024	Migraine	Erenumab	Headache Impact Test (HIT-6)	-6.97
38785393	24.05.2024	Chronic Migraine	Erenumab	HIT-6 total score	-9.5 [0.6]
37940860	09.11.2023	Migraine	Erenumab	HIT-6 score	68.0 (63.8-73.3), 60.0 (54.0-65.0), ...
33421995	11.01.2021	Chronic Migraine	Erenumab	HIT-6 score	p=0.001

Tabelle 7-2: Alte Records für MIDAS Scores für das Medikament Erenumab

PubMed ID	Publikationsdatum	Krankheit	Medikament	Medical Outcome	Result
38785393	24.05.2024	Chronic Migraine	Erenumab	MIDAS score	-22.1 [1.2]
36856538	02.03.2023	Migraine	Anti-CGRP-R mAbs (Erenumab)	MIDAS scores	Comparable to anti-CGRP group at T3 and T6
37254581	31.05.2023	Migraine	Erenumab	Reduction of MIDAS score	-29.2 (-23.0, -35.4)

7.2.7 Anmerkungen zu den Records

Es gibt in dieser DB vielfältige Daten für viele Medikamente, Outcomes, Ergebnisse und Krankheiten. So können in der DB analog für verschiedene Medikamente, Therapien, Outcomes und Krankheiten Abfragen und Analysen ausgeführt werden, um die notwendigen Insights aus den Daten und Paper zu generieren. Falls die Daten nicht sinnvoll erscheinen oder mehr Informationen benötigt werden, kann über die PubMed ID und den Titel das dazugehörige Paper gefunden und gelesen werden.

In dieser Vorstellung der Ergebnisse wurden nur wenige Beispiele für mögliche Abfragen und Analysen gezeigt. Sie dienen als Grundlage und Inspiration für weitere Abfragen mit anderen Daten. Die Daten des Projekts liegen in einem öffentlichen GitHub-Repository frei zur Verfügung als SQL-Dump- und CSV-Datei im Ordner sql_dumps. Der Link zum Repository ist: <https://github.com/1Zero64/PubMedMA>. Gerne können die Daten auf einer lokalen eigenen DB geklont werden, um mit diesen zu experimentieren und/oder zu forschen. Weitere Tabellen und Daten werden im Anhang D präsentiert.

7.3 Bewertung der Datenbank

Die in der PubMed-DB erfassten Daten bieten eine reiche Quelle an Informationen und Wissen, die für verschiedene Analysen genutzt werden können. Durch die detaillierte Aufzeichnung von Therapie- und Medikationsformen lässt sich nachvollziehen, welche Medikamente bei welchen Krankheiten eingesetzt wurden und welche Wirkungen sie gezeigt haben. Die DB enthält zudem teilweise Informationen darüber, auf welche Patientengruppen sich die einzelnen Behandlungen beziehen, was eine gezielte Analyse von Subgruppen ermöglicht.

Die Kategorisierung der Daten erlaubt es, die Auswirkungen der Medikamente auf spezifische medizinische Bereiche zu bestimmen und deren Ausmaß zu quantifizieren. Diese strukturierte Darstellung ist besonders wertvoll für Meta-Analysen, da sie die systematische Zusammenführung und Auswertung von Daten aus verschiedenen Studien ermöglicht. Dadurch können umfassendere und robustere Schlussfolgerungen gezogen werden, die über die Ergebnisse einzelner Studien hinausgehen.

Im Kontext von Meta-Analysen bietet die DB die Möglichkeit, unterschiedliche Therapieansätze zu vergleichen und deren Effektivität und Sicherheit zu bewerten. Durch die Aggregation von Daten aus einer Vielzahl von Quellen können Muster und Trends identifiziert werden, die in kleineren, isolierten Studien möglicherweise nicht erkennbar sind. Dies trägt dazu bei, evidenzbasierte Entscheidungen in der klinischen Praxis zu unterstützen und die medizinische Versorgung zu verbessern.

Als Übersicht eignet sich die DB sehr gut, da sie einen Überblick gibt welches Paper welche Krankheiten und Medikamente behandeln. So können interessierte Personen und Unternehmen den Namen der Paper erhalten und diese selbst auswerten.

Auch können damit Trends erkannt werden, über das Zählen der Paper für ein Medikament nach Publikationsjahr. So kann erkannt werden, welche Medikamente gerade im Kommen (sie werden über die Zeit populärer) und am Gehen (es gibt weniger oder kritischere Tests darüber) sind.

Zusammenfassend bietet die DB durch ihre detaillierte und vielfältige Erfassung von Therapien und Medikamenten eine wertvolle Grundlage für Meta-Analysen und systematische Übersichtsarbeiten. Sie ermöglicht es, umfassende und differenzierte Einblicke in die Wirksamkeit und Anwendung von medizinischen Behandlungen zu gewinnen, was die Basis für fundierte wissenschaftliche und klinische Entscheidungen bildet.

Jedoch ist zu erwähnen, dass die Daten derzeit noch eingeschränkt sind und hauptsächlich Migräne betrachten. In Zukunft müssen hier weitere Krankheiten und passende Medikamente mit aufgenommen werden. Auch ist die Vielfalt der Daten und ihrer Repräsentation ein Hindernis für gute Analyse. Die Medical Outcomes müssen soweit es geht standardisiert werden, um so Analysen zu vereinfachen. Auch Medikamente und Krankheiten brauchen eine Vereinheitlichung, da verschiedene Repräsentationen vorhanden sind. Besonders bei Medikamenten ist es sichtbar, da hier neben dem Namen noch die Dosis angegeben wird. Bei Krankheiten werden ab und zu die Abkürzungen angegeben.

Besonders Medical Outcome, Results und ganz besonders Medication Group variieren stark bei den verschiedenen Läufen. Hier hat LLaMa3 die meisten Probleme beim Erkennen.

8 Fazit und Ausblick

Abschließend lässt sich sagen, dass das Forschungsprojekt erfolgreich umgesetzt wurde. Mit PubMedMA konnte eine Datenbank (DB) mit Medical Records erstellt werden, die viele Einblicke und Analysen ermöglicht und zahlreiche Daten für verschiedene Krankheiten, Medikamente und Therapien liefert. Durch die Kategorisierung der Medical Outcomes und das konsequente Einhalten dieser Kategorien können die Medical Records einfach miteinander verglichen werden. So lässt sich bestimmen, ob ein Outcome durch ein Medikament generell zu einer Verbesserung oder Verschlechterung geführt hat.

Die Accuracy der Daten kann zwar nicht genau bestimmt werden, jedoch ist zu erwarten, dass die Dunkelziffer der Genauigkeit höher ist. Die Evaluation zeigt bereits, dass die generierten Daten eine hohe Ähnlichkeit mit den echten Daten aufweisen und vermutlich identisch sind. Eine stichprobenartige manuelle Überprüfung bestätigte dies. Nichtsdestotrotz wird ein erweitertes Konzept benötigt, um eine bessere und genauere Evaluation der Analyse und Ergebnisse durch LLaMa3 zu ermöglichen.

Mit einem Algorithmus und dynamisch gut konstruierten Prompts konnte eine Pipeline zur Analyse entwickelt werden, die automatisiert und stabil durchläuft und ein Post-Processing in Form von SQL-Queries zum Persistieren der Daten inkludiert. Mit kurzen Analyseläufen und den Ergebnissen als Feedback wurde der Analyse-Prozess iterativ verbessert. Insbesondere die Prompts konnten nach und nach verbessert werden, wodurch qualitativ hochwertigere Records generiert wurden. Jedoch wird eine lange Zeit für die Analyse weniger Paper benötigt. Trotzdem zeigt dieses Projekt, dass mithilfe von LLM komplexe NLP-Aufgaben integrativ und automatisiert bearbeitet werden können.

Das Forschungsprojekt hat mit LLaMa3 viele hochwertige Medical Records generiert, die die Grundlage für die PubMedMA-DB bilden. Diese DB bietet detaillierte Einblicke in verschiedene Krankheiten, Medikamente und Therapien und ist nützlich für die Übersicht über wissenschaftliche Paper und aktuelle Trends. Die strukturierte und umfassende Datenerfassung erleichtert das Verständnis medizinischer Zusammenhänge und die Identifikation von Mustern in der Forschung. Zwar liegt der Fokus derzeit hauptsächlich auf Migräne, was die Nutzung der DB limitiert, jedoch soll sich dies in naher Zukunft ändern, sodass weitere Krankheiten und Medikamente abgedeckt werden.

Besonders hervorzuheben ist die Nutzung der PubMedMA-DB für Meta-Analysen. Die systematische Kategorisierung und die vielfältigen Records erlauben es, Daten aus verschiedenen Quellen zusammenzuführen und fundierte, belastbare Schlussfolgerungen zu ziehen. Dies macht die DB zu einem potenziell wertvollen Werkzeug für die evidenzbasierte Medizin und die Verbesserung der Patientenversorgung in der Zukunft. Weiter kann sie Unternehmen im medizinischen Bereich dabei unterstützen richtige Entscheidungen bei der Entwicklung von Medikamenten, Investition und Strategie zu treffen.

Insgesamt stellt die PubMedMA-DB eine bedeutende Ressource für die medizinische Forschung dar, die kontinuierlich erweitert und verbessert wird, um den steigenden Anforderungen gerecht zu werden und noch tiefere Einblicke in die medizinische Praxis zu ermöglichen. Sie wird zwar nicht als primäre Quelle für die Forschung und Entwicklung von Therapien genutzt, eignet sich jedoch als einer der ersten Schritte für die Sammlung von Daten, Vergleichen und Ergebnissen. Zusätzlich dient sie als Sprungbrett zu einer großen Menge an Papers und medizinischen Studien.

In einer zukünftigen Weiterentwicklung kann die Logik weiter verbessert werden, um die Effizienz und Qualität der Records zu erhöhen. Mit einem weiteren Architekturansatz wird ein Prozess zur Optimierung der Ausgabe verfolgt, bei dem Feedback in die Antworten einfließt. So kann eine zweite Anfrage für dasselbe Paper ausgeführt werden, inklusive der schon generierten Daten. In diesem neuen Prompt wird das LLM gefragt, ob diese Medical Records im Kontext des Abstracts sinnvoll sind. Eventuell nimmt LLaMa3 hier Verbesserungen vor. Die Fehlerrate durch nicht valide JSON-Dateien nimmt zusätzlich ab, da LLaMa3 schon ein korrektes oder fast valides JSON als Vorlage bekommt.

Eine weitere Möglichkeit besteht darin, eine zusätzliche, externe DB anzufragen, um so nochmals eine Überprüfung der generierten Medical Records durchzuführen. Beispielsweise könnte eine spezielle medizinische DB genutzt werden, die detaillierte Informationen zu Medikamenten und deren Wirkungen enthält. Diese externen Quellen könnten die generierten Daten validieren und potenzielle Inkonsistenzen aufdecken. Ebenso können sie als zusätzliche Daten für LLaMa3 dienen, entweder im initialen Prompt zum Generieren von mehr Records oder in einem zweiten als Plausibilitätscheck.

Durch diese erweiterten Feedback- und Validierungsmechanismen wird nicht nur die Qualität und Genauigkeit der generierten Records erhöht, sondern auch die Robustheit des Systems. Solche iterativen und integrativen Ansätze tragen dazu bei, die Nutzung von LLM in der medizinischen Forschung und deren Anwendungen weiter zu optimieren, indem sie zuverlässigere und validierte medizinische Informationen liefern.

Auch die Optimierung des jetzigen Zustands in einer zukünftigen Iteration ist möglich und notwendig, wie die Ergebnisse zeigen. Hier kann insbesondere bei der Evaluierung und beim Prompt Engineering weiter verbessert werden. Besonders die Evaluation und Ähnlichkeitsanalyse über BERT bietet Raum für Optimierung, da gleiche Ergebnisse nicht als korrekt angegeben werden, wie die manuellen Analysen gezeigt haben. Ebenso kann der Algorithmus zur Analyse verbessert werden. Dieser benötigt derzeit zwölf Stunden für knapp 1.100 Paper. Mit einer verbesserten und optimierten Pipeline können hier mehrere Stunden eingespart und bessere Ergebnisse geliefert werden. Weiter wird ein Konzept zur Standardisierung der Daten für besseres Gruppieren und Analysieren benötigt. Damit soll die Vielfalt der Repräsentationen in den Records minimiert werden.

Interessant wäre es, das Verhalten und die Ergebnisse eines LLM mit mehr Parametern zu sehen. Hier sind bessere Ergebnisse mit höheren Genauigkeiten zu erwarten, jedoch ist das Ausmaß unbekannt. Zukünftig können die Ergebnisse nur besser werden, da neuere LLM intelligenter und schneller werden durch bessere und effizientere Algorithmen. Dadurch können zukünftig auch kleinere LLM eine bessere Performance bieten als derzeitige LLM in derselben Größenordnung. LLaMa3.1 mit 405 Milliarden Parametern wurde am 23.07.2024 veröffentlicht und frei zur Verfügung gestellt. Eine Analyse damit wäre interessant, da die ersten Tests etwas bessere Ergebnisse liefern als LLaMa3. Jedoch ist mit der jetzigen Umgebung eine Pipeline mit diesem Modell undenkbar. Einige Prompts wurden im Anhang D.3 mit LLaMa3.1 getestet und mit LLaMa3 verglichen.

Es gibt viele Möglichkeiten, dieses Projekt in der Zukunft iterativ zu verbessern, um die DB zu erweitern und den Nutzen für Meta-Analysen zu maximieren. Die kontinuierliche Verbesserung und Erweiterung der Datenbank wird es ermöglichen, noch umfassendere und genauere Analysen durchzuführen und letztlich zur Verbesserung der medizinischen Forschung und Patientenversorgung beizutragen.

A. Data Collection

Ein Paper, Abstract bzw. Zusammenfassung in PubMed kann online über die Webseite der NCBI gesucht werden. Ein Paper ist dabei oft wie folgt aufgebaut:

> [Brain Inj.](#) 2023 Jan 13:1-10. doi: 10.1080/02699052.2023.2165156



Online ahead of print.

Tolerability of pharmacological agents in the treatment of headache following brain injury: a scoping review

Heather M MacKenzie [1](#) [2](#), Michael Robinson [3](#), Amanda McIntyre [4](#)

Affiliations + expand

PMID: 36637191



DOI: [10.1080/02699052.2023.2165156](https://doi.org/10.1080/02699052.2023.2165156)



Abstract

Background: While systematic reviews have examined medication effectiveness for post-traumatic headache (PTH), they have not assessed tolerability.

Objective: To conduct a scoping review to characterize the adverse effects of pharmacotherapy for PTH.

Methods: CINAHL, CMA Infobase, Cochrane Library, Embase, Epistemonikos, MEDLINE, PEDro, PsycInfo, Scopus, SportDiscus, TRIP and the University of York Center for Reviews and Dissemination were searched. Studies meeting these criteria were included 1) English language, 2) involved humans with traumatic brain injury (TBI), 3) a medication for PTH was administered and 4) reported tolerability outcomes. Author(s), publication year, country of origin, study design, sample demographics, medication type, comparator, dose, treatment duration, adverse effect type and rate, discontinuation rate, and effectiveness outcomes were extracted.

Results: The search yielded 2941 records; 11 studies were included (n = 324 subjects). All subjects had mild TBI except for one with moderate TBI. The following therapies were examined 1) abortive (dihydroergotamine N = 1; metoclopramide N = 1; indomethacin N = 3), 2) prophylactic (divalproex sodium N = 1; amantadine N = 1; erenumab N = 2; amitriptyline N = 2). No serious adverse effects occurred. Observed adverse effects overlap with common symptoms of TBI.

Conclusion: The unique needs of people with TBI must be considered when instituting pharmacotherapy. More studies specifically evaluating medication tolerability in PTH are needed.

Keywords: Drug-Related side effects and adverse reactions; Post-Traumatic headache; Traumatic brain injuries; drug therapy.

[PubMed Disclaimer](#)

Abbildung A-1: Aufbau eines Papers bzw. Abstracts in PubMed

Bei diesem Beispiel handelt es sich um das Paper <https://pubmed.ncbi.nlm.nih.gov/36637191/>.

Results, Conclusion und Objective sind dabei optional, jedoch in den meisten Fällen dabei. Conclusion eher weniger.

Wie vorher erwähnt gibt PyMed nicht das gesamte Resultat zurück und schneidet mittendrin ab. Deswegen wird metapub verwendet, da diese das Resultat immer vollständig wieder gibt. Nachfolgend ist ein Beispiel für dieses Problem zu sehen. Dabei handelt es sich um das Paper <https://pubmed.ncbi.nlm.nih.gov/34743579/> (siehe Abbildung A-2):

PyMed: Seven hundred and seventy-seven patients were randomised (from 22 February 2019 to 29 July, 2020) and 95.1% completed the study. In the erenumab group, 10.6% discontinued medication due to adverse events compared to 38.9% in the topiramate group (odds ratio, 0.19; 95% confidence interval 0.13-0.27;

metapub: Seven hundred and seventy-seven patients were randomised (from 22 February 2019 to 29 July, 2020) and 95.1% completed the study. In the erenumab group, 10.6% discontinued medication due to adverse events compared to 38.9% in the topiramate group (odds ratio, 0.19; 95% confidence interval 0.13-0.27; $p < 0.001$). Significantly more patients achieved a $\geq 50\%$ reduction in monthly migraine days from baseline with erenumab (55.4% vs. 31.2%; odds ratio 2.76; 95% confidence interval 2.06-3.71; $p < 0.001$). No new safety signals occurred.

Leider wurde keine Antwort gefunden, wieso PyMed im Vergleich zu metapub nicht alle Daten liefert. Eine Vermutung liegt auf der falschen Codierung von "5% confidence interval 0.13-0.27; $p < 0.001$ ", da PyMed hier mittendrinnen einen Cut hat und hier mathematische Zeichen codiert sind.

Erenumab versus topiramate for the prevention of migraine – a randomised, double-blind, active-controlled phase 4 trial

Uwe Reuter ^{1 2}, Marc Ehrlich ³, Astrid Gendolla ⁴, Axel Heinze ⁵, Jan Klatt ⁶, Shihua Wen ⁷,
Peggy Hours-Zesiger ⁶, Jacqueline Nickisch ³, Christian Sieder ³, Christian Hentschke ³,
Monika Maier-Peusel ³

Affiliations + expand

PMID: 34743579 PMCID: PMC8793299 DOI: 10.1177/03331024211053571

Abstract

Background: We compared the tolerability and efficacy of erenumab, a monoclonal antibody binding to the calcitonin gene-related peptide receptor, to topiramate for migraine prophylaxis in adults.

Methods: HER-MES was a 24-week, randomised, double-blind, double-dummy, controlled trial conducted in 82 sites in Germany. Patients with ≥ 4 migraine days per month and naïve to study drugs were randomly assigned (1:1) to either subcutaneous erenumab (70 or 140 mg/month) plus topiramate placebo (erenumab group) or oral topiramate at the individual dose with optimal efficacy (50-100 mg/day) plus erenumab placebo (topiramate group). The primary endpoint was medication discontinuation due to an adverse event during the double-blind phase. The proportion of patients that achieved $\geq 50\%$ reduction from baseline in monthly migraine days during the last 3 months of the double-blind phase was a secondary endpoint.

Results: Seven hundred and seventy-seven patients were randomised (from 22 February 2019 to 29 July, 2020) and 95.1% completed the study. In the erenumab group, 10.6% discontinued medication due to adverse events compared to 38.9% in the topiramate group (odds ratio, 0.19; 95% confidence interval 0.13-0.27; $p < 0.001$). Significantly more patients achieved a $\geq 50\%$ reduction in monthly migraine days from baseline with erenumab (55.4% vs. 31.2%; odds ratio 2.76; 95% confidence interval 2.06-3.71; $p < 0.001$). No new safety signals occurred.

Conclusions: Erenumab demonstrated a favourable tolerability and efficacy profile compared to topiramate. Trial registration: ClinicalTrials.gov NCT03828539, URL: <https://clinicaltrials.gov/ct2/show/NCT03828539>.

Keywords: CGRP; Erenumab; head-to-head study; migraine; prophylaxis; topiramate.

Abbildung A-2: Echte Daten des Papers mit der PubMed ID 34743579

Die nachfolgenden Bilder beschreiben den master-DF und zeigen den Aufbau mit ein paar Paper-Daten (siehe Abbildung A-4). In Abbildung Abbildung A-3 werden die Datentypen der einzelnen Attribute beschrieben. Das Attribut medicatio_group_tested ist ein Boolean und repräsentiert die durch LLaMa3 gesetzte Flag. Shape gibt die Dimensionen des DF wieder. Dabei gibt es 2.548 Zeilen, die jeweils ein Paper repräsentieren, und 9 Spalten, die die Attribute repräsentieren. Die Gruppierung nach der Flag zeigt, dass 1.102 True sind und 1.446 False. Die 1.102 True-Flag Paper werden für die Analyse verwendet.

```
Shape: (2548, 9)
```

```
Size: 22932
```

```
Medication Groups tested: medication_group_tested
False    1446
True     1102
dtype: int64
```

```
DTypes: pubmed_id          object
papername            object
authors              object
publication_date     object
abstract             object
results              object
conclusions          object
journal              object
medication_group_tested   bool
dtype: object
```

Abbildung A-3: Beschreibung des master-DF

	pubmed_id	papername \
0	39030528	Efficacy and safety of atogepant, a small mole...
1	39030417	A Retrospective Analysis of Disease Epidemiolo...
2	39028222	Evaluating the Phenotypic Patterns of Post-Tra...
3	39027895	Analyzing Google Search Trends for Migraine Su...
4	39024586	Use of medications for migraine in Aotearoa Ne...
5	39023425	Hyperactivity of the medial thalamus in patient...
6	39023397	Migraine epidemiology in collegiate student-at...
7	39023388	Migraine heritability and beyond: A scoping re...
8	39022731	Interictal widespread pressure hyperalgesia an...
9	39021751	Open Label Pilot of Lorcaserin (a serotonin 2C...
	authors publication_date \	
0	Hou, Min; Luo, Xiaofeng; He, Shuangshuang; Yan...	2024-07-20
1	Terruzzi, Alessandro; AlMadani, Abubaker; Al-R...	2024-07-20
2	Lyons, Hannah S; Sassani, Matilde; Thaller, Ma...	2024-07-19
3	Bishay, Anthony E; Fijany, Arman J; Holan, Col...	2024-07-19
4	Inlach, Fiona; Garrett, Sue	2024-07-18
5	Suzuki, Yukihisa; Kiyosawa, Motohiro; Wakakura...	2024-07-18
6	Burns, Karlee N; Zhao, Huaqing; Master, Christ...	2024-07-18
7	Olfsson, Isa Amalie	2024-07-18
8	Toriyama, Toshihide; Hanaoka, Yoshiki; Horiuchi...	2024-07-18
9	Brezing, Christina A; Sibai, Mohammad; Choi, C...	2024-07-18
	abstract \	
0	BACKGROUND: Migraine is one of the most common...	
1	INTRODUCTION: Migraine is a recurrent, disabli...	
2	INTRODUCTION: Mild traumatic brain injury (TBI...	
3	BACKGROUND: Nurtec, a versatile migraine medic...	
4	AIM: To document and assess acute and preventi...	
5	OBJECTIVE: To examine cerebral functional alte...	
6	OBJECTIVE: This cross-sectional study evaluate...	
7	OBJECTIVE: This scoping review aimed to summar...	
8	BACKGROUND: Patients with vestibular migraine ...	
9	BACKGROUND AND OBJECTIVE: Cannabis Use Disorde...	
	results \	
0	4 RCTs with a total of 2813 subjects met our i...	
1	The study included 203,222 patients (mean age:...	
2	Eight papers related to the military populatio...	
3	For "Nurtec," a significant surge in RSV occur...	
4	Most respondents had used simple analgesics fo...	
5	Compared with the healthy group, both the migr...	
6	Migraine was reported in 5.6% (2617/47,060; 95...	
7	The search identified 52 twin studies on migra...	
8	The study included 163 patients: 31 with VM, 5...	
9	17 participants enrolled, and 14 received medi...	
	conclusions \	
0	This meta-analysis suggests that atogepant is ...	
1	None	
2	None	
3	The analysis of RSV trends for "Nurtec" and "m...	
4	None	
5	The medial thalamus may be associated with the...	
6	None	
7	None	
8	None	
9	Lorcaserin was well tolerated in males but not...	
	journal	medication_group_tested
0	The journal of headache and pain	True
1	Pain and therapy	False
2	Military medicine	False
3	Plastic and reconstructive surgery. Global open	False
4	The New Zealand medical journal	True
5	Headache	False
6	Headache	False
7	Headache	False
8	Frontiers in neurology	False
9	Journal of substance use	True

Abbildung A-4: Aufbau des master-DF mit Daten von einigen Paper

B. Prompt Engineering

In diesem Anhang werden einige dynamisch aufgebaute Prompts und die Antworten von LLaMa3 vorgestellt.

B.1 Filter Prompt

Ausgabe der Analyse für das Paper mit der PubMed ID 39030528 und dem Titel „Efficacy and safety of atogepant, a small molecule CGRP receptor antagonist, for the preventive treatment of migraine: a systematic review and meta-analysis.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies and medicaments for future meta analysis. Read the following content and answer if any group, medication or therapy groups are tested and results are given in this paper. Just return yes or no.

BACKGROUND: Migraine is one of the most common diseases worldwide while current treatment options are not ideal. New therapeutic classes of migraine, the calcitonin gene-related peptide (CGRP) antagonists, have been developed and shown considerable effectiveness and safety. The present study aimed to systematically evaluate the efficacy and safety of atogepant, a CGRP antagonist, for migraine prophylaxis from the results of randomized controlled trials (RCTs).

METHODS: The Cochrane Library, Embase, PubMed and <https://www.>

CLINICALTRIALS: gov/ were searched for RCTs that compared atogepant with placebo for migraine prophylaxis from inception of the databases to Feb 1, 2024. Outcome data involving efficacy and safety were combined and analyzed using Review Manager Software version 5.3 (RevMan 5.3). For each outcome, risk ratios (RRs) or standardized mean difference (SMD) were calculated.

RESULTS: 4 RCTs with a total of 2813 subjects met our inclusion criteria. The overall effect estimate showed that atogepant was significantly superior to placebo in terms of the reduction of monthly migraine (SMD -0.40, 95% CI -0.46 to -0.34) or headache (SMD -0.39, 95% CI -0.46 to -0.33) days, the reduction of acute medication use days (SMD -0.45, 95% CI -0.51 to -0.39) and 50% responder rate (RR 1.66, 95% CI 1.46 to 1.89), while no dose-related improvements were found between different dosage groups. For the safety, significant number of patients experienced treatment-emergent adverse events (TEAEs) with atogepant than with placebo (RR 1.10, 95% CI 1.02-1.21) while there was no obvious difference between the five dosage groups. Most TEAEs involved constipation (RR 2.55, 95% CI 1.91-3.41), nausea (RR 2.19, 95% CI 1.67-2.87) and urinary tract infection (RR 1.49, 95% CI 1.05-2.11). In addition, a high dosage of atogepant may also increase the risk of treatment-related TEAEs (RR 1.64, 95% CI 1.02-2.63) and fatigue (RR 3.07, 95% CI 1.13-8.35).

CONCLUSIONS: This meta-analysis suggests that atogepant is effective and tolerable for migraine prophylaxis including episodic or chronic migraine compared with placebo. It is critical to weigh the benefits of different doses against the risk of adverse events in clinical application of atogepant. Longer and multi-dose trials with larger sample sizes are required to verify the current findings.

Output:

Yes.

B.2 Group Prompt

Ausgabe der Analyse für das Paper mit der PubMed ID 38965463 und dem Titel „Persistent effectiveness of CGRP antibody therapy in migraine and comorbid medication overuse or medication overuse headache - a retrospective real-world analysis.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medicaments for future meta-analysis. Extract the number of tested groups for a medical test from the following data. Just return the number of groups as an integer without other text.

BACKGROUND: Management of patients with migraine who have concomitant medication overuse (MO) or medication overuse headache (MOH) is a major problem in clinical practice. Detoxification of acute analgesics before or during initiation of prophylactic therapy has long been recommended although this concept has recently been questioned. Additionally, relapse after detoxification is a common problem. This real-world study analyses the initial and sustained effectiveness of prophylactic migraine therapy with CGRP (receptor) antibodies without prior detoxification in patients with comorbid MO or MOH for up to one year.

METHODS: A retrospective real-world analysis was performed on 291 patients (episodic migraine (EM) with MO (EM-MO; n=35), EM without MO (EM-noMO; n=77), chronic migraine (CM) with MOH (CM-MOH; n=109), CM without MOH (CM-noMOH; n=70). All patients began treatment with either erenumab (n=173), fremanezumab (n=70) or galcanezumab (n=48) without prior detoxification. Data were available for up to 12 months of treatment. Responder rates for monthly headache days (MHD), monthly migraine days (MMD) and monthly acute medication intake (AMD) were analysed.

RESULTS: All groups showed a significant reduction in MHD, MMD and AMD at the last observed time point compared to baseline. In patients with CM and MOH, 60.6% (66/109) no longer fulfilled the definition of MO or MOH and a further 13.8% (15/109) had only EM-MO. In the EM cohort, 89% (31/35) of MO patients lost their MO during therapy. MHD and AMD 30% responder rates were comparable for CM-MOH and CM-noMOH (MHD: CM-MOH: 56.0% vs. CM-noMOH: 41.4%, p=0.058, AMD: CM-MOH: 66.1% vs. CM-noMOH: 52.9%, p=0.077). MMD responder rate did not differ significantly (after Bonferroni adjustment) (CM-MOH: 62.4% vs. CM-noMOH: 47.1%, p=0.045, $\alpha=0.017$). After successful initiation of therapy, 15.4% of the initial CM-MOH patients relapsed and met the criterion for CM-MOH at the end of follow-up. There were no antibody specific differences in response to therapy.

CONCLUSIONS: Our data confirms the effectiveness of CGRP antibody treatment in migraine patients with additional MOH or MO in a real-world setting. Low relapse rates after initial successful therapy support an early start of CGRP antibody treatment in patients with MOH or MO.

TRIAL REGISTRATION: No registration, retrospective analysis.

Output:

4

Ausgabe der Analyse für das Paper mit der PubMed ID 38901375 und dem Titel „Efficacy and safety of venlafaxine versus nortriptyline for the preventive treatment of migraine: A double-blind randomized clinical trial.”

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medicaments for future meta-analysis. Extract the number of tested groups for a medical test from the following data. Just return the number of groups as an integer without other text.

OBJECTIVE: Migraine, as a primary headache disorder, stands as one of the primary causes of disability worldwide. Consequently, prophylactic treatments are highly recommended for individuals experiencing recurrent migraine episodes. Our study aimed to compare the efficacy and safety profiles of venlafaxine and nortriptyline in the prophylactic management of migraine.

METHODS: In this single-center, randomized, double-blind clinical trial, 210 migraine patients were allocated into two groups in a 1:1 ratio. One group received venlafaxine (37.5mg, orally twice daily), while the other group administered nortriptyline (25mg, orally once daily). A neurologist documented (1) headache intensity using the Visual Analog Scale (VAS) and 6-point Behavioral Rating Scale (BRS-6), (2) headache frequency (per month), and (3) headache duration (in hours) of participants on days 0, 45, and 90 of the intervention.

RESULTS: Following the 90-day intervention, a significant decrease was observed in VAS, BRS-6, frequency, and duration of headaches within both groups (all with p-values <0.001). No difference in VAS, BRS-6, or headache durations was observed between the two groups after 45 and 90 days of treatment (all p-values > 0.05). Although the headache frequency exhibited no difference between the groups after 45 days (p-value = 0.097), a significantly lower frequency in the venlafaxine group was observed at day 90 of the intervention (p-value = 0.011). The reductions in attack parameters in the 0-45- and 0-90-day intervals did not meet statistical significance between the two groups (p-values > 0.05). 77.0% of the participants in the venlafaxine group and 79.2% in the nortriptyline group experienced a minimum of 50 % improvement in all attack parameters. Venlafaxine demonstrated a statistically significant lower incidence of adverse reactions in comparison to nortriptyline (p-value = 0.005). A total of 33 adverse drug reactions were documented in the venlafaxine group and 53 in the nortriptyline group, with insomnia observed in the former and xerostomia in the latter as the most prevalent side effects.

CONCLUSIONS: Venlafaxine and nortriptyline demonstrate clinically significant and comparable therapeutic efficacy for migraine patients in reducing the intensity, frequency, and duration of headache attacks. Venlafaxine may be preferred to nortriptyline in the context of migraine preventive treatment under comparable conditions due to its lower incidence of adverse effects.

Output:

2

Oft entspricht die Anzahl der Medication Groups der Anzahl an generierten Medical Records. Für manche Gruppen werden ab und zu auch mehrere Records generiert wie bei diesem Paper. Die Ausgabe der Records für dieses Paper ist in B.3 zu finden. Ausgabe der Analyse für das Paper mit der PubMed ID 38898657 und dem Titel „A 3-year follow-up study of outcomes associated with patterns of traditional acute and preventive migraine treatment: An administrative claims-based cohort study in the United States.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the number of tested groups for a medical test from the following data. Just return the number of groups as an integer without other text.

OBJECTIVE: To describe treatment patterns and direct healthcare costs over 3 years following the initiation of standard-of-care acute and preventive migraine medications in patients with migraine in the United States.

BACKGROUND: There are limited data on long-term (>1 year) migraine treatment patterns and associated outcomes.

METHODS: This was a retrospective, observational cohort study using US claims data from the IBM® MarketScan® Research Database (January 2010–December 2017). Adults were included if they had a prescription claim for acute migraine treatments (AMT) or preventive migraine treatments (PMT) in the index period (January 2011–December 2014). The AMT cohort was categorized as persistent, cycled, or added-on subgroups; the PMT cohort was categorized as PMT-persistent, switched without gaps, or cycled with gaps. Migraine-specific annual direct costs (2017 US\$) across AMT and PMT cohort subgroups were summarized at baseline through 3 years from index (follow-up).

RESULTS: During the index period, 20,778 and 42,259 patients initiated an AMT and a PMT, respectively. At the 3-year follow-up, migraine-specific direct costs were lower in the persistent subgroup relative to the non-persistent subgroups in both AMT (mean [SD]: \$789 [\$1741] vs. \$2847 [\$8149] in the added-on subgroup and \$862 [\$5426] for the cycled subgroup) and PMT cohorts (mean [SD]: \$1817 [\$5892] in the persistent subgroup vs. \$4257 [\$11,392] in the switched without gaps subgroup and \$3269 [\$18,540] in the cycled with gaps subgroup). Acute medication overuse was lower in the persistent subgroup (1025/6504 [27.2%]) vs. non-persistent subgroups (11,236/58,863 [32.2%] in cycled with gaps subgroup and 1431/6504 [39.4%] in the switched without gaps subgroup). Most patients used multiple acute (19,717/20,778 [94.9%]) or preventive (38,494/42,259 [91.1%]) pharmacological therapies over 3 years following treatment initiation. Gaps in preventive therapy were common; an average gap ranged from 85 to 211 days (~3–7 months).

CONCLUSION: Migraine-specific annual healthcare costs and acute migraine medication overuse remained lowest among patients with persistent AMT and PMT versus non-persistent treatment. Study findings are limited to the US population. Future studies should compare costs and associated outcomes between newer preventive migraine medications in patients with migraine.

Output:

2

Selten wird eine hohe Anzahl an Gruppen ausgegeben, die nicht der Anzahl an Records entspricht. In einem Fall wurden sogar 13 bei einem anderen 131 Gruppen beobachtet, wobei nur zwei Records generiert wurden für eine Gruppe. Ausgabe der Analyse für das Paper mit der PubMed ID 38729773 dem Titel „Non-vascular ATP-sensitive potassium channel activation does not trigger migraine attacks: A randomized clinical trial.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the number of tested groups for a medical test from the following data. Just return the number of groups as an integer without other text.

OBJECTIVE: To investigate the role of NN414, a selective K_{ATP} channel opener for the Kir6.2/SUR1 channel subtype found in neurons and β -pancreatic cells, in inducing migraine attacks in individuals with migraine without aura.

METHODS: Thirteen participants were randomly allocated to receive NN414 and placebo on two days separated by at least one week. The primary endpoint was the difference in the incidence of migraine attacks after NN414 compared with placebo. The secondary endpoints were the difference in the area under the curve for headache intensity scores, middle cerebral artery blood flow velocity (V_{MCA}), superficial temporal artery diameter, heart rate, and mean arterial pressure.

RESULTS: Twelve participants completed the study, with two (16.6%) reporting migraine attacks after NN414 compared to one (8.3%) after placebo ($p=0.53$). The area under the curve for headache intensity, V_{MCA}, superficial temporal artery diameter, heart rate, and mean arterial pressure did not differ between NN414 and placebo ($p>0.05$, all comparisons).

CONCLUSION: The lack of migraine induction upon activation of the Kir6.2/SUR1 channel subtype suggests it may not contribute to migraine pathogenesis. Our findings point to K_{ATP} channel blockers that target the Kir6.1/SUR2B subtype, found in cerebral vasculature, as potential candidates for innovative antimigraine treatments. Registration number: NCT04744129.

Output:

13

B.3 Records Prompt

Die Antworten zu den Prompts zur Generierung von Medical Records beinhalten neben der Liste weiteren Text. Unter anderem wird hier argumentiert, wieso diese Records so erstellt wurden und wieso Daten übersprungen oder leere Listen geliefert wurden, falls dies passiert ist

Ausgabe der Analyse für das Paper mit der PubMed ID 34743579 und dem Titel „Erenumab versus topiramate for the prevention of migraine - a randomised, double-blind, active-controlled phase 4 trial.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medicaments for future meta-analysis.

Extract the disease, tested group or patients, therapy (e.g. chemotherapy) or medicament or program, medical outcome, and results from the given information.

Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present:

disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record, skip it and do not return it.

If no records are found, return an empty list.

Result should only have the value with unit of measurement. Answers like '31.2% reduction' for result are not allowed.

If no tested_group_or_patients is specified, return 'No group specified'.

Here is an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

BACKGROUND: We compared the tolerability and efficacy of erenumab, a monoclonal antibody binding to the calcitonin gene-related peptide receptor, to topiramate for migraine prophylaxis in adults.

METHODS: HER-MES was a 24-week, randomised, double-blind, double-dummy, controlled trial conducted in 82 sites in Germany. Patients with ≥4 migraine days per month and naïve to study drugs were randomly assigned (1:1)

to either subcutaneous erenumab (70 or 140 mg/month) plus topiramate placebo (erenumab group) or oral topiramate at the individual dose with optimal efficacy (50-100 mg/day) plus erenumab placebo (topiramate group). The primary endpoint was medication discontinuation due to an adverse event during the double-blind phase. The proportion of patients that achieved ≥50% reduction from baseline in monthly migraine days during the last 3 months of the double-blind phase was a secondary endpoint.

RESULTS: Seven hundred and seventy-seven patients were randomised (from 22 February 2019 to 29 July 2020) and 95.1% completed the study. In the erenumab group, 10.6% discontinued medication due to adverse events compared to 38.9% in the topiramate group (odds ratio, 0.19; 95% confidence interval 0.13-0.27; $p < 0.001$). Significantly more patients achieved a ≥50% reduction in monthly migraine days from baseline with erenumab (55.4% vs. 31.2%; odds ratio 2.76; 95% confidence interval 2.06-3.71; $p < 0.001$). No new safety signals occurred.

CONCLUSIONS: Erenumab demonstrated a favourable tolerability and efficacy profile compared to topiramate. Trial registration: ClinicalTrials.gov NCT03828539, URL: <https://clinicaltrials.gov/ct2/show/NCT03828539>.

Output:

Here is the extracted list of medical records in JSON format:

```
[  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Randomised group",  
    "therapy_or_medication": "Erenumab",  
    "medical_outcome": "discontinuation due to adverse events",  
    "result": "10.6%"  
  }, {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Randomised group",  
    "therapy_or_medication": "Erenumab",  
    "medical_outcome": "≥50% reduction from baseline in monthly migraine days",  
    "result": "55.4%"  
  }, {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Randomised group",  
    "therapy_or_medication": "Topiramate",  
    "medical_outcome": "discontinuation due to adverse events",  
    "result": "38.9%"  
  }, {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Randomised group",  
    "therapy_or_medication": "Topiramate",  
    "medical_outcome": "discontinuation due to adverse events",  
    "result": "31.2%"  
  }]
```

```

    "medical_outcome": ">50% reduction from baseline in monthly migraine days",
    "result": "31.2%"
}
]

```

Note that I skipped the record with the trial registration information, as it did not meet the criteria for disease, therapy/medication, tested group/patients, medical outcome, or result.

Auch erkennt LLaMa3 ob Zahlen für Resultate angegeben werden oder nicht. Ausgabe der Analyse für das Paper mit der PubMed ID 38700141 und dem Titel „The impact of the migraine treatment onabotulinumtoxinA on inflammatory and pain responses: Insights from an animal model.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the disease, tested group or responder rates or problem groups or patients, therapy (e.g. chemotherapy) or medication or program, medical outcome, and results from the given information. Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present: disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result.

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record skip it and do not return it. If no records are found, return an empty list. Result should only have the value with the unit of measurement. Answers like '31.2% reduction' for result are not allowed. If no tested_group_or_patients is specified return 'No group specified'.

Here is an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, Male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2% Migraine, Group with MMD, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

OBJECTIVE: Migraine, a prevalent and debilitating disease, involves complex pathophysiology possibly including inflammation and heightened pain sensitivity. The current study utilized the complete Freund's adjuvant (CFA)

model of inflammation, with onabotulinumtoxinA (BoNT/A) as a treatment of interest due to its use in clinical migraine management. Using an animal model, the study sought to investigate the role of BoNT/A in modulating CFA-induced inflammation, alterations in pain sensitivity, and the regulation of calcitonin gene-related peptide (CGRP) release. Further, we aimed to assess the changes in SNAP-25 through western blot analysis to gain insights into the mechanistic action of BoNT/A.

METHODS: BoNT/A or control was administered subcutaneously at the periorbital region of rats 3 days before the induction of inflammation using CFA. Periorbital mechanical sensitivity was assessed post-inflammation, and alterations in CGRP release were evaluated. Changes in SNAP-25 levels were determined using western blot analysis.

RESULTS: Upon CFA-induced inflammation, there was a marked increase in periorbital mechanical sensitivity, with the inflammation side showing increased sensitivity compared to other periorbital areas. BoNT/A did decrease the withdrawal thresholds in the electronic von Frey test. Despite not being able to observe differences in pain thresholds or CGRP release, BoNT/A reduced baseline release under CFA inflamed conditions. Analysis of SNAP-25 levels in the trigeminal ganglion revealed both intact and cleaved forms that were notably elevated in BoNT/A-treated animals. These findings, derived from western blot analysis, suggest an effect on neurotransmitter release.

CONCLUSION: Our investigation highlights the role of BoNT/A in reducing baseline CGRP in the context of inflammation and its involvement in SNAP-25 cleavage. In contrast, BoNT/A did not appear to alter facial pain sensitivity induced by inflammation, suggesting that mechanisms other than baseline CGRP could be implicated in the elevated thresholds in the CFA model.

Output:

Here is the extracted JSON list of records:

```
[  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Rats with periorbital inflammation",  
    "therapy_or_medication": "OnabotulinumtoxinA (BoNT/A)",  
    "medical_outcome": "Decrease in withdrawal thresholds",  
    "result": "Not specified"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Rats with periorbital inflammation",  
    "therapy_or_medication": "OnabotulinumtoxinA (BoNT/A)",  
    "medical_outcome": "Reduction in baseline CGRP release",  
    "result": "Not specified"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Rats with periorbital inflammation",  
    "therapy_or_medication": "OnabotulinumtoxinA (BoNT/A)",  
    "medical_outcome": "Increase in pain sensitivity",  
    "result": "Not specified"  
  }]
```

```

        "therapy_or_medication": "OnabotulinumtoxinA (BoNT/A)",
        "medical_outcome": "Changes in SNAP-25 levels",
        "result": "Not specified"
    }
]

```

Note that there are no specific numbers or percentages mentioned in the provided text, so I couldn't extract any numerical results.

Ausgabe der Analyse für das Paper mit der PubMed ID 38617394 dem Titel „Efficacy of Probiotics in Prevention of Migraine Attacks in Children: A Randomized Clinical Trial Study.”

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the disease, tested group or responder rates or problem groups or patients, therapy (e.g., chemotherapy) or medication or program, medical outcome, and results from the given information. Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present: disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result.

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record, skip it and do not return it. If no records are found, return an empty list. Result should only have the value with the unit of measurement. Answers like '31.2% reduction' for result are not allowed. If no tested_group_or_patients is specified, return 'No group specified'.

Here is an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, Male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2%

Migraine, Group with MMD, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

OBJECTIVES: Migraine is a chronic and joint disease in children. The results of previous studies on the effectiveness of probiotics in preventing migraine attacks in children have been controversial. This study aims to investigate the effect of probiotics on migraine prophylaxis in children.

MATERIALS & METHODS: In this clinical trial study, 41 children aged 5 to 15 with migraine enrolled the study in two control and intervention groups. Children in the intervention group (18 children) received propranolol at a dose of 1 mg per kilogram of body weight daily in two divided doses along with a 250 mg Yomogi capsule daily for three months, and children in the control group (23 children), received propranolol along with placebo for three months. The study compared the frequency and duration of headache days, PedMIDAS criteria, and parental satisfaction between the two groups before treatment, as well as one month and three months post-treatment.

RESULTS: The number of headache days in both groups decreased over time, but in the intervention group, this decrease was more than the control group was statistically significant ($P=0.045$). The average PedMIDAS scale after treatment in the intervention group was 3.9 ± 3.8 ; in the control group, it was 8.4 ± 8.2 , which was statistically significant ($P=0.047$). Parents' satisfaction with the treatment was statistically significantly higher in the intervention group (94.4%) than in the control group (54.5%) ($P=0.011$). No significant drug complications were seen in any of the two groups.

CONCLUSION: In children with migraine, adding probiotics to migraine treatment reduces the intensity and number of days of children's headaches and increases the parents' satisfaction with the treatment.

Output:

Here are the extracted medical records in JSON format:

```
[  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Children with Migraine",  
    "therapy_or_medication": "Propranolol + Probiotics (Yomogi capsule)",  
    "medical_outcome": "Number of headache days",  
    "result": "Not specified"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Children with Migraine",  
    "therapy_or_medication": "Propranolol + Placebo",  
    "medical_outcome": "Number of headache days",  
    "result": "Not specified"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Children with Migraine",  
    "therapy_or_medication": "Propranolol + Probiotics (Yomogi capsule)",  
    "medical_outcome": "Number of headache days",  
    "result": "Not specified"  
  }]
```

```

    "medical_outcome": "PedMIDAS scale",
    "result": "3.9 ± 3.8"
  },
  {
    "disease": "Migraine",
    "tested_group_or_patients": "Children with Migraine",
    "therapy_or_medication": "Propranolol + Placebo",
    "medical_outcome": "PedMIDAS scale",
    "result": "8.4 ± 8.2"
  },
  {
    "disease": "Migraine",
    "tested_group_or_patients": "Children with Migraine",
    "therapy_or_medication": "Propranolol + Probiotics (Yomogi capsule)",
    "medical_outcome": "Parental satisfaction",
    "result": "94.4%"
  },
  {
    "disease": "Migraine",
    "tested_group_or_patients": "Children with Migraine",
    "therapy_or_medication": "Propranolol + Placebo",
    "medical_outcome": "Parental satisfaction",
    "result": "54.5%"
  }
]

```

Note: I extracted multiple records from the paper, each representing a different outcome or result.

Ausgabe der Analyse für das Paper mit der PubMed ID 38898657 dem Titel „A 3-year follow-up study of outcomes associated with patterns of traditional acute and preventive migraine treatment: An administrative claims-based cohort study in the United States.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the disease, tested group or responder rates or problem groups or patients, therapy (e.g., chemotherapy) or medication or program, medical outcome, and results from the given information. Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present: disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result.

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record, skip it and do not return it. If no records are found, return an empty list. Result should only have the value with the unit of measurement. Answers like '31.2% reduction' for result are not allowed. If no tested_group_or_patients is specified, return 'No group specified'.

Here is an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, Male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2%

Migraine, Group with MMD, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

OBJECTIVE: To describe treatment patterns and direct healthcare costs over 3 years following the initiation of standard-of-care acute and preventive migraine medications in patients with migraine in the United States.

BACKGROUND: There are limited data on long-term (>1 year) migraine treatment patterns and associated outcomes.

METHODS: This was a retrospective, observational cohort study using US claims data from the IBM® MarketScan® Research Database (January 2010–December 2017). Adults were included if they had a prescription claim for acute migraine treatments (AMT) or preventive migraine treatments (PMT) in the index period (January 2011–December 2014). The AMT cohort was categorized as persistent, cycled, or added-on subgroups; the PMT cohort was categorized PMT-persistent, switched without gaps, or cycled with gaps. Migraine-specific annual direct costs (2017 US\$) across AMT and PMT cohort subgroups were summarized at baseline through 3 years from index (follow-up).

RESULTS: During the index period, 20,778 and 42,259 patients initiated an AMT and a PMT, respectively. At the 3-year follow-up, migraine-specific direct costs were lower in the persistent subgroup relative to the non-persistent subgroups in both AMT (mean [SD]: \$789 [\$1741] vs. \$2847 [\$8149] in the added-on subgroup and \$862 [\$5426] for the cycled subgroup) and PMT cohorts (mean [SD]: \$1817 [\$5892] in the persistent subgroup vs. \$4257 [\$11,392] in the switched without gaps subgroup and \$3269 [\$18,540] in the

cycled with gaps subgroup). Acute medication overuse was lower in the persistent subgroup (1025/6504 [27.2%]) vs. non-persistent subgroups (11,236/58,863 [32.2%] in cycled with gaps subgroup and 1431/6504 [39.4%] in the switched without gaps subgroup). Most patients used multiple acute (19,717/20,778 [94.9%]) or preventive (38,494/42,259 [91.1%]) pharmacological therapies over 3 years following treatment initiation. Gaps in preventive therapy were common; an average gap ranged from 85 to 211 days (~3-7 months).

CONCLUSION: Migraine-specific annual healthcare costs and acute migraine medication overuse remained lowest among patients with persistent AMT and PMT versus non-persistent treatment. Study findings are limited to the US population. Future studies should compare costs and associated outcomes between newer preventive migraine medications in patients with migraine.

Output:

Here is the extracted JSON list of records:

```
[  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Persistent subgroup",  
    "therapy_or_medication": "Acute Migraine Treatments (AMT)",  
    "medical_outcome": "migraine-specific annual direct costs",  
    "result": "$789"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Non-persistent subgroups (Added-on and Cycled subgroups)",  
    "therapy_or_medication": "Acute Migraine Treatments (AMT)",  
    "medical_outcome": "migraine-specific annual direct costs",  
    "result": "$2847"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Persistent subgroup",  
    "therapy_or_medication": "Preventive Migraine Treatments (PMT)",  
    "medical_outcome": "migraine-specific annual direct costs",  
    "result": "$1817"  
  },
```

```

{
  "disease": "Migraine",
  "tested_group_or_patients": "Non-persistent subgroups (Switched without gaps and Cycled with gaps subgroups)",
  "therapy_or_medication": "Preventive Migraine Treatments (PMT)",
  "medical_outcome": "migraine-specific annual direct costs",
  "result": "$4257"
},
{
  "disease": "Migraine",
  "tested_group_or_patients": "Persistent subgroup",
  "therapy_or_medication": "Acute Migraine Treatments (AMT)",
  "medical_outcome": "acute migraine medication overuse",
  "result": "27.2%"
},
{
  "disease": "Migraine",
  "tested_group_or_patients": "Non-persistent subgroups (Cycled with gaps and Switched without gaps subgroups)",
  "therapy_or_medication": "Acute Migraine Treatments (AMT)",
  "medical_outcome": "acute migraine medication overuse",
  "result": "32.2%"
}
]

```

Note that I did not extract any records related to \"therapy\" or \"medication\" as the text does not specify any particular treatment.

B.4 Evaluation Prompt

Ähnlich wie bei Records argumentiert LLaMa3 basierend auf dem Record und den Results wieso diese Kategorien für Evaluation und Category für das Outcome gewählt wurden. Das hilft bei der manuellen Überprüfung und beim Nachvollziehen der Ergebnisse. Weitestgehend hält sich LLaMa3 an die vorgegebenen Aufzählungen, selten werden neue erfunden oder verschiedene Variationen der bestehenden kreiert (wie in dieser Antwort bei Outcome Evaluation)

Ausgabe der Analyse für das Paper mit der PubMed ID 34743579 und dem Titel „Erenumab versus topiramate for the prevention of migraine - a randomised, double-blind, active-controlled phase 4 trial.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medicaments for future meta-analysis.

You have to determine the category and evaluation of the outcome based on the given information.

Give me a JSON document with the following information: outcome_category, outcome_evaluation.

outcome_category can have only one of these categories:

- Efficacy Measures (such as responder rates)
- Symptom Intensity and Relief (such as pain intensity)
- Adverse Events and Side Effects
- Quality of Life (such as MIDAS score)
- Physiological Measure (such as heart rate)
- Medication Usage (such as number of days on medication)
- Health Economics (such as Cost-effectiveness)
- No Category

Do not create other categories!

outcome_evaluation can have only one of these categories:

- significantly improved
- no significant difference
- significantly worse

Do not create other categories!

Extract the document based on the following information:

```
{'disease': 'Migraine', 'tested_group_or_patients': 'Randomised group',  
'therapy_or_medication': 'Topiramate', 'medical_outcome': '≥50% reduction  
from baseline in monthly migraine days', 'result': '31.2%'}
```

Seven hundred and seventy-seven patients were randomised (from 22 February 2019 to 29 July 2020) and 95.1% completed the study. In the erenumab group, 10.6% discontinued medication due to adverse events compared to 38.9% in the topiramate group (odds ratio, 0.19; 95% confidence interval 0.13-0.27; $p < 0.001$). Significantly more patients achieved a $\geq 50\%$ reduction in monthly migraine days from baseline with erenumab (55.4% vs. 31.2%; odds ratio 2.76; 95% confidence interval 2.06-3.71; $p < 0.001$). No new safety signals occurred.

Output:

Here is the extracted JSON document:

```
{  
    "outcome_category": "Efficacy Measures",  
    "outcome_evaluation": "significantly better"  
}
```

Note: I categorized the outcome as "Efficacy Measures" because it measures the effectiveness of a therapy (topiramate) in reducing the frequency of migraine days, and specifically mentions a $\geq 50\%$ reduction from baseline. The evaluation is considered "significantly better" because the result states that more patients achieved this threshold with topiramate compared to erenumab.

C. Analyse

Falls keine Medical Records gefunden wurden lieferte LLaMa3 eine leere Liste zurück sowie eine Begründung, wieso keine gefunden wurde (siehe Abbildung C-1). Oft gab es Probleme die Records aus den Edge-Cases zu ziehe, da diese übertrieben waren und unrealistische Ergebnisse hatten. Ein Beispiel hierfür die „survival rate of 101%“. Dies ist nicht möglich und LLaMa3 hat dies erkannt und argumentiert, dass es sich hierbei um eine Übertreibung handelt.

```
Patients undergoing chemotherapy exhibited a reported survival rate of 101%, reflecting an apparent error or hyperbolic expression within the data set.  
on 29: [] (empty list)
```

```
No specific disease, therapy/medicament, medical outcome, and results were mentioned in the given text, so no records can be extracted. The provided information only refers to a survival rate of patients undergoing chemo therapy, which is likely an error or hyperbole.
```

Abbildung C-1: Begründungen von LLaMa3 für das Nichtfinden von Records in einem synthetischen Paper

Ein Durchlauf mit einem kombinierten Records und Evaluation Prompt hat zwar schlechtere Ergebnisse geliefert, jedoch war die Verarbeitungszeit mit 4 Stunden und 30 Minuten viel geringer als bei getrennten Prompts (siehe Abbildung Abbildung C-2). Grund dafür ist das iterative dynamische Prompt Building und die Analyse für jeden gefundenen Record. So werden statt nur einem Run für ein Paper mit 20 Sekunden mehrere Runs ausgeführt: Ein Run der 15 Sekunden dauert (da nur der Record Prompt analysiert wird) plus n-Runs, die im Schnitt 10 Sekunden benötigen. n ist hierbei die Anzahl der gefundenen Records. So dauert ein Paper mit sechs Records $15 + 6 * 10 = 75$ Sekunden, statt nur 20 Sekunden.

```
[REDACTED] | 1102/1102 [100%] in 4:25:45.1 (0.07/s  
1102 runs finished.  
2024-07-21 18:29:41.835506 : Analysis done  
Execution time: 04:25:45  
  
Failed runs:  
[]
```

Abbildung C-2: Kurzer Durchlauf mit einem kombinierten Prompt. Dieser dauerte nur knapp 4 Stunden 30 Minuten statt 12 Stunden

Abbildung C-3 zeigt wie wenig Fehler in einem Komplett durchlauf auftreten. Dabei muss die Anzahl der dort sichtbaren Fehler halbiert werden, da ein Fehler zu zwei Exceptions führt durch das Bubble-Up im Exception-Handling. So gab es 30 Fehler, die jedoch aufgefangen und wiederholt wurden. Dabei gab es Paper, die mehrere Fehler verursacht haben, wie zum Beispiel 1006. Hier gab es drei Fehler bei vier Versuchen. Der vierte Versuch funktionierte dann. Kein Paper hatte über fünf Fehler, weshalb es auch keine fehlerhaften Paper gab. Distinkte Fehler gab es 23, somit gab es nur Komplikationen bei 2,1 % der Paper. Die Quote der fehlerhaften Paper liegt bei 0 %.

```

on 29: Fehlerhafter JSON-String:
on 29: There was an error: Expecting ',' delimiter: line 7 column 19 (char 363)
on 33: Fehlerhafter JSON-String:
on 33: There was an error: Expecting ',' delimiter: line 7 column 20 (char 252)
on 49: Fehlerhafter JSON-String:
on 49: There was an error: Expecting ',' delimiter: line 21 column 21 (char 727)
on 58: Fehlerhafter JSON-String:
on 58: There was an error: Extra data: line 9 column 1 (char 253)
on 70: Fehlerhafter JSON-String:
on 70: There was an error: Expecting ',' delimiter: line 7 column 18 (char 283)
on 70: Fehlerhafter JSON-String:
on 70: There was an error: Expecting ',' delimiter: line 7 column 18 (char 263)
on 126: Fehlerhafter JSON-String:
on 126: There was an error: Expecting ':' delimiter: line 69 column 22 (char 3053)
on 242: Fehlerhafter JSON-String:
on 242: There was an error: Expecting ',' delimiter: line 7 column 20 (char 331)
on 379: Fehlerhafter JSON-String:
on 379: There was an error: Expecting ',' delimiter: line 7 column 19 (char 347)
on 398: Fehlerhafter JSON-String:
on 398: There was an error: Expecting ',' delimiter: line 7 column 18 (char 260)
on 416: Fehlerhafter JSON-String:
on 416: There was an error: Expecting ',' delimiter: line 7 column 18 (char 250)
on 446: Fehlerhafter JSON-String:
on 446: There was an error: Expecting ',' delimiter: line 28 column 19 (char 1024)
on 482: Fehlerhafter JSON-String:
on 482: There was an error: Expecting value: line 35 column 15 (char 1087)
on 578: Fehlerhafter JSON-String:
on 578: There was an error: Expecting ',' delimiter: line 7 column 19 (char 268)
on 632: Fehlerhafter JSON-String:
on 632: There was an error: Expecting ',' delimiter: line 42 column 27 (char 1462)
on 673: Fehlerhafter JSON-String:
on 673: There was an error: Expecting ',' delimiter: line 7 column 36 (char 317)
on 709: Fehlerhafter JSON-String:
on 709: There was an error: Expecting ',' delimiter: line 14 column 19 (char 463)
on 711: There was an error: 'outcome_evaluation'
on 937: Fehlerhafter JSON-String:
on 937: There was an error: Expecting ',' delimiter: line 7 column 19 (char 281)
on 950: Fehlerhafter JSON-String:
on 950: There was an error: Expecting ',' delimiter: line 14 column 23 (char 377)
on 950: Fehlerhafter JSON-String:
on 950: There was an error: Expecting ',' delimiter: line 14 column 23 (char 361)
on 969: Fehlerhafter JSON-String:
on 969: There was an error: Expecting ',' delimiter: line 7 column 17 (char 314)
on 1006: Fehlerhafter JSON-String:
on 1006: There was an error: Expecting ',' delimiter: line 7 column 19 (char 307)
on 1006: Fehlerhafter JSON-String:
on 1006: There was an error: Expecting ',' delimiter: line 7 column 19 (char 306)
on 1006: Fehlerhafter JSON-String:
on 1006: There was an error: Expecting ',' delimiter: line 7 column 19 (char 307)
on 1034: Fehlerhafter JSON-String:
on 1034: There was an error: Expecting ',' delimiter: line 14 column 20 (char 604)
on 1035: Fehlerhafter JSON-String:
on 1035: There was an error: Expecting ',' delimiter: line 7 column 19 (char 287)
on 1084: Fehlerhafter JSON-String:
on 1084: There was an error: Expecting ',' delimiter: line 14 column 33 (char 572)

```

Abbildung C-3: Stabilität der Pipeline mit wenigen Fehlern

D. Ergebnisse und Diskussion

D.1 Alte Daten

Nachfolgend ist eine Übersicht der generierten Medical Records zu sehen von einem alten Durchlauf mit einer alten Struktur. Hier sind Medication Group, Outcome Category sowie Evaluation nicht dabei. In dieser View wurden auch weniger Paper ausgegeben, da Zeilen mit einer fehlenden Krankheit, fehlendem Medical Outcome und Result rausgefiltert wurden. So wurde die Anzahl der Records von knapp 3.400 Zeilen auf knapp 2.700 reduziert. Diese Übersicht soll die Vielfalt der Daten zeigen und wie gut LLaMa3 Result und Medical Outcome extrahieren konnte.

Auch zeigt es Probleme der alten Methodik auf. Bei der Behandlung einer Krankheit mit dem gleichen Medikament bei verschiedenen Gruppen, kann es vorkommen, dass die Resultate des gleichen Medical Outcomes bei den Gruppen gemessen wird. In dem Fall gibt es Zeilen, die das gleiche Medikament und das gleiche Medical Outcome betrachten, jedoch mit unterschiedlichen Werten. Dies ist verwirrend, da die Resultate nicht zugeordnet werden können. Mit der Angabe der Medication Group, die in der alten Version fehlte, wird dies klar.

Id [PK] integer	pubmed_Id integer	paper character varying (255)	text character varying (255)	authors character varying (2)	journal character varying (2)	publication_date date	disease character varying (255)	therapy_or_medicament character varying (255)	medical_outcome character varying (255)	result character varying (255)	
2383	3350	39001504	Systemic Treatment of...	Himmelman, Ve...	Cancers	2024-07-13	Hepatocellular carcin...	Sorafenib	Progression free survival	5.7 months	
2384	3351	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TAF	HBeAg loss	22.4%	
2385	3352	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TDF	HBeAg loss	11.2%	
2386	3353	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TAF	Antiviral response	65%	
2387	3354	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TDF	Antiviral response	54.5%	
2388	3355	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TAF	ALT normalization	93.9%	
2389	3356	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TDF	ALT normalization	81.2%	
2390	3357	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TAF	Renal function variation	-0.04 to -0.01	
2391	3358	38994860	Clinical Efficacy And S...	Vo, Thong Duy P...	Clinical and tran...	2024-07-12	Chronic HBV	TDF	Renal function variation	0.03 mg/dL every 6 months	
2392	3359	38994860	Impact of baseline hep...	Pan, Di Liu, Hao...	World journal of ...	2024-07-12	Hepatocellular carcin...	Camrelizumab, Lenvatinib	Progression-free survival (PFS)	8.32 months	
2393	3360	38994860	Impact of baseline hep...	Pan, Di Liu, Hao...	World journal of ...	2024-07-12	Hepatocellular carcin...	Camrelizumab, Lenvatinib	Overall survival (OS)	13.30 months	
2394	3361	38994860	Impact of baseline hep...	Pan, Di Liu, Hao...	World journal of ...	2024-07-12	Hepatocellular carcin...	Camrelizumab, Lenvatinib	Objective response rate (ORR)	19.70%	
2395	3474	38945038	Evidence-based interve...	Lin, Jierong; Wan,...	Sleep medicine	2024-07-01	Post-thoracic surger...	Music	pain intensity	0.91; probability: 54%	
2396	3362	38994860	Impact of baseline hep...	Pan, Di Liu, Hao...	World journal of ...	2024-07-12	Hepatocellular carcin...	Camrelizumab, Lenvatinib	Objective response rate (ORR)	33.33%	
2397	3363	38994860	Impact of baseline hep...	Pan, Di Liu, Hao...	World journal of ...	2024-07-12	Hepatocellular carcin...	Camrelizumab, Lenvatinib	Disease control rate (DCR)	72.73%	
2398	3364	38993510	Real-world Effectiveness...	Ji, Fanpu; Tran, S...	Journal of clinic...	2024-07-12	HCV	DAA	SVR12	96.9%	
2399	3365	38984849	Assessment of hepatiti...	Wang, Chih-Wen; ..	Liver Internation...	2024-07-10	Hepatitis B virus	Baseline HBV DNA	Clinical relapse	1.48 (95% CI 1.10-1.99; p < .009)	

Abbildung D-1: Übersicht der Daten in einer alten Version der DB

Die folgende Tabelle soll diese Szenario veranschaulichen:

Tabelle D-1: Zeilen mit schwer verstehbaren Daten

Paper	Disease	Medicament	Medical Outcome	Result
1234	Migraine	Erenumab	Reduction MMD	20 %
1234	Migraine	Erenumab	Reduction MMD	28 %
1234	Migraine	Erenumab	Reduction MMD	40 %

Hier ist nicht klar, was die Zahlen bedeuten und welches Resultat nun richtig ist. Mit der Medication Group wird die Geschichte dieser Records klarer:

Tabelle D-2: Zeile mit leicht verstehbaren Daten

Paper	Disease	Medication Group	Medicament	Medical Outcome	Result
1234	Migraine	Female group	Erenumab	Reduction MMD	20 %
1234	Migraine	Men group	Erenumab	Reduction MMD	28 %
1234	Migraine	Children group	Erenumab	Reduction MMD	40 %

Die verschiedenen Resultate sind nun klar, da sie einer Gruppe zugeordnet werden können. In der DB kann damit zusätzlich auch nach bestimmten Patientengruppen gesucht werden, falls eine bestimmte Gruppe behandelt werden muss.

Die neue Pipeline im Projekt sorgte nun auch dafür, dass es keine Records mehr gibt, die einen der Felder leer haben. Es gibt keine Leerstellen mehr.

D.2 Weitere Tabellen und Daten

In diesem Teil des Anhangs werden Tabellen und Daten präsentiert, die eventuell interessant sein könnten.

Einige Medical Records haben kein richtiges Resultat (wie hier nur 0) oder Fließtext als Resultat. Für diese Records wurde das dazugehörige Paper betrachtet. Bei diesem Paper wurde klar, dass es keine numerischen Resultate zu extrahieren gibt sondern das bei allen Records es keine serious Adverse Effects gab. Deswegen hat LLaMa3 hier 0 gesetzt. Die Records wurden somit gut erkannt. Die Medication Group wurde jedoch nicht erkannt (Subjects with mild TBI and one subject). LLaMa3 hat besonders bei diesem Attribut viele Schwierigkeiten. Dieses Beispiel zeigt, dass nicht alle Paper numerische Resultate haben.

Post-traumatic headache (PTH)	No group specified	Dihydroergotamine	Adverse effects	0	Adverse Events and Side Effects	no significant difference
Post-traumatic headache (PTH)	No group specified	Metoclopramide	Adverse effects	0	Adverse Events and Side Effects	no significant difference
Post-traumatic headache (PTH)	No group specified	Indomethacin	Adverse effects	0	Adverse Events and Side Effects	no significant difference
Post-traumatic headache (PTH)	No group specified	Divalproex sodium	Prophylactic	0	Adverse Events and Side Effects	no significant difference
Post-traumatic headache (PTH)	No group specified	Amantadine	Prophylactic	0	Adverse Events and Side Effects	no significant difference
Post-traumatic headache (PTH)	No group specified	Erenumab	Prophylactic	0	Adverse Events and Side Effects	no significant difference
Post-traumatic headache (PTH)	No group specified	Amitriptyline	Prophylactic	0	Adverse Events and Side Effects	no significant difference

Abbildung D-2: Records ohne richtiges Resultat

Tolerability of pharmacological agents in the treatment of headache following brain injury: a scoping review

Heather M MacKenzie ^{1 2}, Michael Robinson ³, Amanda McIntyre ⁴

Affiliations + expand

PMID: 36637191 DOI: 10.1080/02699052.2023.2165156

Abstract

Background: While systematic reviews have examined medication effectiveness for post-traumatic headache (PTH), they have not assessed tolerability.

Objective: To conduct a scoping review to characterize the adverse effects of pharmacotherapy for PTH.

Methods: CINAHL, CMA Infobase, Cochrane Library, Embase, Epistemonikos, MEDLINE, PEDro, PsycInfo, Scopus, SportDiscus, TRIP and the University of York Center for Reviews and Dissemination were searched. Studies meeting these criteria were included 1) English language, 2) involved humans with traumatic brain injury (TBI), 3) a medication for PTH was administered and 4) reported tolerability outcomes. Author(s), publication year, country of origin, study design, sample demographics, medication type, comparator, dose, treatment duration, adverse effect type and rate, discontinuation rate, and effectiveness outcomes were extracted.

Results: The search yielded 2941 records; 11 studies were included (n = 324 subjects). All subjects had mild TBI except for one with moderate TBI. The following therapies were examined 1) abortive (dihydroergotamine N = 1; metoclopramide N = 1; indomethacin N = 3), 2) prophylactic (divalproex sodium N = 1; amantadine N = 1; erenumab N = 2; amitriptyline N = 2). No serious adverse effects occurred. Observed adverse effects overlap with common symptoms of TBI.

Conclusion: The unique needs of people with TBI must be considered when instituting pharmacotherapy. More studies specifically evaluating medication tolerability in PTH are needed.

Keywords: Drug-Related side effects and adverse reactions; Post-Traumatic headache; Traumatic brain injuries; drug therapy.

Abbildung D-3: Paper ohne numerische Resultate

Die Anzahl der veröffentlichten Paper pro Jahr wird in der folgenden Tabelle dargestellt. Es ist ein signifikanter Anstieg an Paper von 2020 bis 2024 zu sehen. Mehr als die Hälfte der 1.102 analysierten Paper wurden in 2024 veröffentlicht.

year numeric	paper_count bigint
2024	616
2023	162
2022	68
2021	57
2020	26
2019	24
2018	8
2017	9
2016	9
2015	14
2014	16
2013	12
2012	5
2011	4
2010	3
2009	1

Abbildung D-4: Anzahl der veröffentlichten Paper pro Jahr

Die nachfolgende Abbildung zeigt die Anzahl der Records gruppiert nach Outcome Category und Evaluation. Hier ist sichtbar, das Efficacy Measures mit significantly improved am meisten vorkommt mit 1.351 Records. Danach gibt es einen großen Abfall zu 276 Records für Symptom Intensity and Relief. Die meisten Records haben significantly improved. Worse kam weniger häufig vor. Sind die Daten alle korrekt klassifiziert, kann davon ausgegangen werden, dass die meisten Studien erfolgreich verliefen.

outcome_category character varying (255)	outcome_evaluation character varying (255)	count bigint
Efficacy Measures	significantly improved	1351
Symptom Intensity and Relief	significantly improved	276
Physiological Measure	significantly improved	244
Adverse Events and Side Effects	significantly worse	210
Efficacy Measures	no significant difference	206
Efficacy Measures	significantly worse	184
Physiological Measure	significantly worse	154
Physiological Measure	no significant difference	123
Adverse Events and Side Effects	no significant difference	115
Quality of Life	significantly improved	110
Efficacy Measures	significantly better	75
Symptom Intensity and Relief	no significant difference	55
Medication Usage	no significant difference	54
No Category	no significant difference	46
Adverse Events and Side Effects	significantly improved	44
Medication Usage	significantly worse	41
Medication Usage	significantly improved	35
Symptom Intensity and Relief	significantly worse	29
Efficacy Measures	significantly improved	29
Health Economics	significantly improved	23
Health Economics	significantly worse	11
Physiological Measure	significantly reduced	10
Health Economics	significantly better	9
Physiological Measure	significantly lower	9

rows: 138 of 138 | Query complete 00:00:00.166

Abbildung D-5: Anzahl der Records nach Outcome Category und Evaluation

D.3 Vergleich LLaMa3.1 und LLaMa3

Vor wenigen Tagen wurde am 23.07.2024 LLaMa3.1 released, das neueste Modell von Meta mit 405 Milliarden Parameter. Eine Umsetzung der Analyse mit diesem Model war nicht möglich, da die Hardware-Ressourcen bei Weitem nicht ausreichen.

Dennoch wurde einige Tests ausgeführt über eine Bereitstellung von LLaMa3.1 im Internet. In diesen Tests wurden LLaMa3.1 die gleichen Prompts gegeben, wie für LLaMa3. Die Antworten der beiden sind nachfolgend aufgeschrieben.

Ausgabe der Analyse für das Paper mit der PubMed ID 34743579 und dem Titel „Erenumab versus topiramate for the prevention of migraine - a randomised, double-blind, active-controlled phase 4 trial.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medicaments for future meta-analysis.

Extract the disease, tested group or patients, therapy (e.g. chemotherapy) or medicament or program, medical outcome, and results from the given information.

Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present:

disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record, skip it and do not return it.

If no records are found, return an empty list.

Result should only have the value with unit of measurement. Answers like '31.2% reduction' for result are not allowed.

If no tested_group_or_patients is specified, return 'No group specified'.

Here is an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

BACKGROUND: We compared the tolerability and efficacy of erenumab, a monoclonal antibody binding to the calcitonin gene-related peptide receptor, to topiramate for migraine prophylaxis in adults.

METHODS: HER-MES was a 24-week, randomised, double-blind, double-dummy, controlled trial conducted in 82 sites in Germany. Patients with ≥4 migraine days per month and naïve to study drugs were randomly assigned (1:1) to either subcutaneous erenumab (70 or 140 mg/month) plus topiramate placebo (erenumab group) or oral topiramate at the individual dose with optimal efficacy (50-100 mg/day) plus erenumab placebo (topiramate group). The primary endpoint was medication discontinuation due to an adverse event during the double-blind phase. The proportion of patients that achieved ≥50% reduction from baseline in monthly migraine days during the last 3 months of the double-blind phase was a secondary endpoint.

RESULTS: Seven hundred and seventy-seven patients were randomised (from 22 February 2019 to 29 July 2020) and 95.1% completed the study. In the erenumab group, 10.6% discontinued medication due to adverse events compared to 38.9% in the topiramate group (odds ratio, 0.19; 95% confidence interval 0.13-0.27; $p < 0.001$). Significantly more patients achieved a ≥50% reduction in monthly migraine days from baseline with erenumab (55.4% vs. 31.2%; odds ratio 2.76; 95% confidence interval 2.06-3.71; $p < 0.001$). No new safety signals occurred.

CONCLUSIONS: Erenumab demonstrated a favourable tolerability and efficacy profile compared to topiramate. Trial registration: ClinicalTrials.gov NCT03828539, URL: <https://clinicaltrials.gov/ct2/show/NCT03828539>.

Output LLaMa3.1:

Here is the extracted JSON list of records:

```

[

```
{ "disease": "Migraine",
 "tested_group_or_patients": "Erenumab group",
 "therapy_or_medication": "Erenumab",
 "medical_outcome": "discontinuation",
 "result": "10.6%"

}, { "disease": "Migraine",
 "tested_group_or_patients": "Topiramate group",
 "therapy_or_medication": "Topiramate",
 "medical_outcome": "discontinuation",
 "result": "38.9%"

}, { "disease": "Migraine",
 "tested_group_or_patients": "Erenumab group",
 "therapy_or_medication": "Erenumab",
 "medical_outcome": "reduction MMD",
 "result": "55.4%"

}, { "disease": "Migraine",
 "tested_group_or_patients": "Topiramate group",
```

```

 "therapy_or_medication": "Topiramate",
 "medical_outcome": "reduction MMD",
 "result": "31.2%"
}
]
```

```

Note that I extracted four records, two for each group, with the corresponding disease, tested group or patients, therapy or medication, medical outcome, and result. I made sure to keep the results concise and only included the value with the unit of measurement, as per your requirements. Let me know if you need further assistance!

Output LLaMa3:

Here is the extracted list of medical records in JSON format:

```
[
  {
    "disease": "Migraine",
    "tested_group_or_patients": "Randomised group",
    "therapy_or_medication": "Erenumab",
    "medical_outcome": "discontinuation due to adverse events",
    "result": "10.6%"

  }, {
    "disease": "Migraine",
    "tested_group_or_patients": "Randomised group",
    "therapy_or_medication": "Erenumab",
    "medical_outcome": "\u226550% reduction from baseline in monthly migraine days",
    "result": "55.4%"

  }, {
    "disease": "Migraine",
    "tested_group_or_patients": "Randomised group",
    "therapy_or_medication": "Topiramate",
    "medical_outcome": "discontinuation due to adverse events",
    "result": "38.9%"

  }, {
    "disease": "Migraine",
    "tested_group_or_patients": "Randomised group",
    "therapy_or_medication": "Topiramate",
    "medical_outcome": "\u226550% reduction from baseline in monthly migraine days",
    "result": "31.2%"
  }
]
```

```
    }  
]  
  
Note that I skipped the record with the trial registration information, as  
it did not meet the criteria for disease, therapy/medication, tested  
group/patients, medical outcome, or result.
```

Ausgabe der Analyse für das Paper mit der PubMed ID 38946982 und dem Titel „Stroke Risk Reduction in Migraine Patients Using Propranolol: Evidence from Two Large-Scale Real-World Data Analyses.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the disease, tested group or responder rates or problem groups or patients, therapy (e.g., chemotherapy) or medication or program, medical outcome, and results from the given information. Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present: disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result.

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record, skip it and do not return it. If no records are found, return an empty list. Result should only have the value with the unit of measurement. Answers like '31.2% reduction' for result are not allowed. If no tested_group_or_patients is specified, return 'No group specified'.

Here is an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, Male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2%

Migraine, Group with MMD, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

BACKGROUND: Propranolol, a non-selective beta-blocker, is commonly used for migraine prevention, but its impact on stroke risk among migraine patients remains controversial. Using two large electronic health records-based datasets, we examined stroke risk differences between migraine patients with and without documented use of propranolol.

METHODS: This retrospective case-control study utilized EHR data from the Vanderbilt University Medical Center (VUMC) and the All of Us Research

Program. Migraine patients were first identified based on the International Classification of Headache Disorders, 3rd edition (ICHD-3) criteria using diagnosis codes. Among these patients, cases were defined as those with a primary diagnosis of stroke following the first diagnosis of migraine, while controls had no stroke after their first migraine diagnosis. Logistic regression models, adjusted for potential factors associated with stroke risk, assessed the association between propranolol use and stroke risk, stratified by sex and migraine subtype. A Cox proportional hazards regression model was used to estimate the hazard ratio (HR) for stroke risk at 1, 2, 5, and 10 years from baseline.

RESULTS: In the VUMC database, 378 cases and 15,209 controls were identified, while the All of Us database included 267 cases and 6,579 controls. Propranolol significantly reduced stroke risk in female migraine patients (VUMC: OR=0.52, p=0.006; All of Us: OR=0.39, p=0.007), but not in males. The effect was more pronounced for ischemic stroke and in females with migraines without aura (MO) (VUMC: OR=0.60, p=0.014; All of Us: OR=0.28, p=0.006). The Cox model showed lower stroke rates in propranolol-treated female migraine patients at 1, 2, 5, and 10 years (VUMC: HR=0.06-0.55, p=0.0018-0.085; All of Us: HR=0.23, p=0.045 at 10 years).

CONCLUSIONS: Propranolol is associated with a significant reduction in stroke risk, particularly ischemic stroke, among female migraine without aura patients. These findings suggest that propranolol may benefit stroke prevention in high-risk populations.

Output LLaMa3.1:

Here are the extracted records in JSON format:

```
[  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Females with migraines without aura (MO)",  
    "therapy_or_medication": "Propranolol",  
    "medical_outcome": "stroke risk",  
    "result": "0.60 (OR)"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Females with migraines without aura (MO)",  
    "therapy_or_medication": "Propranolol",  
    "medical_outcome": "stroke risk",  
    "result": "0.28 (OR)"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Female migraine patients",  
    "therapy_or_medication": "Propranolol",  
    "medical_outcome": "stroke risk",  
    "result": "0.52 (OR)"  
  }]
```

```

} ,
{
  "disease": "Migraine",
  "tested_group_or_patients": "Female migraine patients",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "stroke risk",
  "result": "0.39 (OR)"

} ,
{
  "disease": "Migraine",
  "tested_group_or_patients": "Female migraine patients",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "stroke rate",
  "result": "0.06-0.55 (HR)"

} ,
{
  "disease": "Migraine",
  "tested_group_or_patients": "Female migraine patients",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "stroke rate",
  "result": "0.23 (HR)"

}
]

```

Note that I did not include the male group as the results mentioned that propranolol did not have a significant effect on stroke risk in males. Also, I did not include the results for the overall population as the text only mentions the results for females with migraines without aura (MO) and female migraine patients. Let me know if you'd like me to extract any further information!

Output LLaMa3:

Here is the extracted JSON list of records:

```
[
  {
    "disease": "Migraine",
    "tested_group_or_patients": "Female migraine patients",
    "therapy_or_medication": "Propranolol",
    "medical_outcome": "Stroke risk (VUMC)",
    "result": "OR=0.52"

} ,
{
  "disease": "Migraine",

```

```

    "tested_group_or_patients": "Female migraine patients",
    "therapy_or_medication": "Propranolol",
    "medical_outcome": "Stroke risk (All of Us)",
    "result": "OR=0.39"
},
{
  "disease": "Migraine",
  "tested_group_or_patients": "Females with migraines without aura (MO)",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "Ischemic stroke risk (VUMC)",
  "result": "OR=0.60"
},
{
  "disease": "Migraine",
  "tested_group_or_patients": "Females with migraines without aura (MO)",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "Ischemic stroke risk (All of Us)",
  "result": "OR=0.28"
},
{
  "disease": "Migraine",
  "tested_group_or_patients": "Female migraine patients",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "Stroke rate at 10 years (VUMC)",
  "result": "HR=0.06-0.55"
},
{
  "disease": "Migraine",
  "tested_group_or_patients": "Female migraine patients",
  "therapy_or_medication": "Propranolol",
  "medical_outcome": "Stroke rate at 10 years (All of Us)",
  "result": "HR=0.23"
}
]

```

Let me know if you'd like me to extract more records or if there's anything else I can help with!

Ausgabe der Analyse für das Paper mit der PubMed ID 38755541 und dem Titel „Effectiveness and tolerability of eptinezumab in treating patients with migraine resistant to conventional preventive medications and CGRP (receptor) antibodies: a multicentre retrospective real-world analysis from Germany.“

Input:

You are an assistant to read medical abstracts and results from papers about diseases, therapies, and medications for future meta-analysis. Extract the disease, tested group or responder rates or problem groups or patients, therapy (e.g. chemotherapy) or medication or program, medical outcome, and results from the given information. Multiple records out of one paper are possible. The extracted answers must be very short.

Give me a JSON list of records with the following attributes, they must be present: disease, tested_group_or_patients, therapy_or_medication, medical_outcome, result.

If disease, therapy_or_medication, tested_group_or_patients, medical_outcome, or result is not specified / not identified for a record skip it and do not return it. If no records are found, return an empty list. Result should only have the value with the unit of measurement. Answers like '31.2% reduction' for result are not allowed. If no tested_group_or_patients is specified, return 'No group specified'.

Here an example of how a medical record must look like:

Migraine, Females with migraines without aura (MO), Erenumab, discontinuation, 10.6%

Migraine, Male patients, Topiramate, discontinuation, 38.9%

Migraine, Randomised group, Erenumab, reduction MMD, 55.4%

Migraine, No group specified, Topiramate, reduction MMD, 31.2%

Migraine, Group with MMD, Topiramate, reduction MMD, 31.2%

Extract the records for 2 groups from the following information in short:

BACKGROUND: Eptinezumab is a monoclonal antibody that targets calcitonin gene-related peptide (CGRP mAb) and is used for migraine prophylaxis. Efficacy data are mainly from clinical trials, real-world data are hardly available yet. Reimbursement policy in Germany leads to eptinezumab mainly being used in patients having failed pre-treatment with other CGRP mAb. To date, it is unclear whether eptinezumab is efficacious and well tolerated in this population and how the treatment response differs from patients who are naive to CGRP mAbs.

METHODS: We analysed clinical routine data of 79 patients (episodic migraine (EM): n=19; chronic migraine (CM): n=60) from four different centres in Germany. All patients were treated with eptinezumab (100mg). Differences in monthly headache (MHD), migraine (MMD) and acute medication days (AMD) after three months were analysed. The correlation of response with the number of CGRP mAb failures was evaluated. Significance level has been corrected ($\alpha=0.017$).

RESULTS: After three months MHD, MMD and AMD were significantly reduced. In EM, the median reduction for MHD was 4.0 days (IQR: -6.5 to -1.0; p=0.001), for MMD 3.0 days (IQR: -5.5 to -1.5; p<0.001) and for AMD 2.0 days (IQR: -5.0 to -0.5; p=0.006). In CM, median reduction of MHD was 4 days (IQR: -8.0 to 0.0; p<0.001), 3.0 days (IQR: -6.0 to -1.0; p<0.001) for MMD and 1.0 day (IQR: -5.0 to 0.0; p<0.001) for AMD. All patients were resistant to conventional preventive therapies and most to CGRP mAbs. Fourteen patients had never received a CGRP mAb and 65 patients had received at least one mAb without sufficient effectiveness and/or intolerance (one: n=20, two: n=28, three: n=17). There was a significant association between the number of prior therapies and the 30% MHD responder rate (none: 78.6%, one: 45.0%, two: 32.1%, three: 23.5%, p=0.010). Regarding tolerability, 10.4% (8/77) reported mild side effects.

CONCLUSIONS: The effectiveness of eptinezumab is significantly reduced in patients who have not previously responded to other CGRP mAbs. However, limitations such as the retrospective nature of the analysis, the small sample size and the short treatment period with only the lower dose of eptinezumab must be considered when interpreting the results.

Output LLaMa3.1:

Here is the JSON list of extracted records:

```
[  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Episodic migraine (EM) patients",  
    "therapy_or_medication": "Eptinezumab",  
    "medical_outcome": "Reduction in monthly headache days (MHD)",  
    "result": "4.0 days"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Episodic migraine (EM) patients",  
    "therapy_or_medication": "Eptinezumab",  
    "medical_outcome": "Reduction in monthly migraine days (MMD)",  
    "result": "3.0 days"  
  },  
  {  
    "disease": "Migraine",  
    "tested_group_or_patients": "Episodic migraine (EM) patients",  
    "therapy_or_medication": "Eptinezumab",  
    "medical_outcome": "Reduction in acute medication days (AMD)",  
    "result": "2.0 days"  
  },  
  {  
    "disease": "Migraine",
```

```

    "tested_group_or_patients": "Chronic migraine (CM) patients",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Reduction in monthly headache days (MHD)",
    "result": "4.0 days"
},
{
    "disease": "Migraine",
    "tested_group_or_patients": "Chronic migraine (CM) patients",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Reduction in monthly migraine days (MMD)",
    "result": "3.0 days"
},
{
    "disease": "Migraine",
    "tested_group_or_patients": "Chronic migraine (CM) patients",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Reduction in acute medication days (AMD)",
    "result": "1.0 day"
},
{
    "disease": "Migraine",
    "tested_group_or_patients": "Patients with no prior CGRP mAb treatment",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "30% MHD responder rate",
    "result": "78.6%"
}
]

```

Output LLaMa3:

Here are the extracted medical records in JSON format:

```
[
{
    "disease": "Migraine",
    "tested_group_or_patients": "Episodic Migraine (EM)",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Monthly Headache (MHD) reduction",
    "result": "-4.0 days"
}, {
    "disease": "Migraine",

```

```

    "tested_group_or_patients": "Episodic Migraine (EM)",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Migraine (MMD) reduction",
    "result": "-3.0 days"
}, {
    "disease": "Migraine",
    "tested_group_or_patients": "Episodic Migraine (EM)",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Acute Medication Days (AMD) reduction",
    "result": "-2.0 days"
}, {
    "disease": "Migraine",
    "tested_group_or_patients": "Chronic Migraine (CM)",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Monthly Headache (MHD) reduction",
    "result": "-4.0 days"
}, {
    "disease": "Migraine",
    "tested_group_or_patients": "Chronic Migraine (CM)",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Migraine (MMD) reduction",
    "result": "-3.0 days"
}, {
    "disease": "Migraine",
    "tested_group_or_patients": "Chronic Migraine (CM)",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "Acute Medication Days (AMD) reduction",
    "result": "-1.0 day"
}, {
    "disease": "Migraine",
    "tested_group_or_patients": "Patients resistant to conventional preventive therapies and most to CGRP mAbs",
    "therapy_or_medication": "Eptinezumab",
    "medical_outcome": "30% MHD responder rate",
    "result": "78.6%"
}
]

```

Note that I skipped the record about tolerability (side effects) as it doesn't meet the specified format.

Die Ergebnisse der beiden Modelle sind sehr gut miteinander vergleichbar. Die Anzahl der Records ist bei beiden Modellen bei allen drei Beispielen identisch.

Für die Studie mit der PubMed ID 34743579 hat LLaMa3.1 die Datensätze mit detaillierten gruppenspezifischen Bezeichnungen wie "Erenumab group" und "Topiramate group" extrahiert, während LLaMa3 eine allgemeinere Bezeichnung "Randomised group" für beide Gruppen verwendet hat, was spezifische Unterschiede verschleiern kann. LLaMa3 hat hier die Gruppen nicht erkannt. LLaMa3.1 hat die bereitgestellten Beispiele direkt nachgebildet, um Klarheit bei den medizinischen Ergebnissen zu gewährleisten. Auch wurden die Antworten kurzgehalten wie gefordert. Andererseits hat LLaMa3 detailliertere Medical Records bereitgestellt und die Prozentwerte einbezogen, was wertvolle Kontextinformationen hinzufügt, aber nicht den Anforderungen an die Kürze entspricht.

Für die Studie mit der PubMed ID 38946982 haben beide Modelle die gleichen spezifischen Gruppen, Resultate und Medical Outcomes erkannt. Beide sind sich sehr ähnlich in den Antworten, wobei LLaMa3 wieder mehr Detailtiefe hat mit zusätzlichem Kontext, der zwar informativ, aber nicht den Extraktionsrichtlinien entspricht.

Für die Studie mit der PubMed ID 38755541 waren die Datensätze bei beiden prägnant und entsprachen dem erwarteten Ausgabeformat. Ein großer Unterschied ist, dass LLaMa3 negative Werte für Reduktionen bereitstellt, die zwar von der Logik her korrekt sind, jedoch zu Verwirrung führen können da eine negative Reduktion als Erhöhung verstanden werden kann. Hier ist LLaMa3.1 stärker mit klar verständlichen Records

Insgesamt zeigte LLaMa3.1 eine stärkere Einhaltung der vorgegebenen Richtlinien für die Extraktion und Formatierung von Daten, was Klarheit und Konsistenz gewährleistete. LLaMa3, obwohl es zusätzlichen Kontext und Detailtiefe bot, wichen manchmal von der strikten Formatierung und Spezifität ab, die für eine präzise Datenerfassung im Kontext einer Meta-Analyse erforderlich ist. Beide Modelle haben weiteren Inhalt neben den Records generiert, obwohl dies verboten wurde im Prompt. Dieses Problem liegt wohl in der LLaMa3 Familie.

Zusammenfassend lässt sich sagen, dass beide Modelle gute Ergebnisse geliefert haben, wobei LLaMa3.1 was Regeln, Spezifität und Verständlichkeit leicht vorne liegt. Bei der Erkennung der Gruppen liegt LLaMa3.1 weiter vorne, da beim ersten Beispiel die Gruppen immer korrekt wiedergegeben wurden. LLaMa3 hatte hier auch bei mehreren Läufen immer Probleme. Dennoch hat LLaMa3 insgesamt bei diesen Tests eine gute Performance geliefert. Mit LLaMa3.1 wurde dies nochmal bestätigt.

Jedoch ist zu erwähnen, dass hier nur wenige Paper analysiert wurden. Diese liefen zwar beide gute Ergebnisse, jedoch ist das kein Indiz dafür dass die Leistung der beiden Modelle bei komplexeren Paper und bei einer viel größeren Anzahl weiterhin vergleichbar bleibt. Es ist zu erwarten, dass hier LLaMa3.1 einen größeren Vorsprung bekommt, da durch die Anzahl der Parameter bessere Ergebnisse bei komplexeren Aufgaben geliefert werden können.